

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет бизнеса и менеджмента

ДОМАШНЕЕ ЗАДАНИЕ №2 ПО КУРСУ
“КЛАССИФИКАЦИЯ СТАТИСТИЧЕСКИХ ДАННЫХ”

Григоращенко Екатерины Андреевны

2 курс, образовательная программа «Бизнес-информатика»

Москва, 2021 год

Оглавление

| | |
|------------------------------------|----|
| Оглавление | 2 |
| 1. Формирование базы данных | 3 |
| 2. Дискриминантный анализ | 8 |
| 3. Построение дерева классификации | 15 |
| 4. Декомпозиция смеси | 25 |
| Источники | 29 |

1. Формирование базы данных

В данном исследовании будет изучено влияние различных признаков на качество в разных странах. В качестве независимых непрерывных переменных были взяты показатели: индекс качества жизни, паритет покупателя, индекс безопасности, индекс здравоохранения, индекс стоимости проживания, соотношение стоимости жилья и заработной платы населения. В качестве бинарных независимых переменных взяты степень загрязнения воздуха (1 - состояние находится в пределах нормы, 0 - состояние не находится в пределах нормы) и климат (1 - более благоприятный климат, 0 - менее благоприятный климат). В качестве целевой (зависимой переменной) был взят ответ на вопрос, находится ли страна в топе 50 лучших стран для проживания или нет. “1” обозначает принадлежность к топу 50, “2” - обратное.

Задача классификация: определить является ли страна более комфортной для проживания людей или нет.

Таблица 1. Исходные данные

| Страна | Индекс качества жизни | Паритет покупательной способности | Индекс безопасности | Индекс с здравоохранения | Индекс стоимости проживания | Соотношения стоимости и жилья и заработка | Степень загрязнения воздуха | Климат | Присутствие в топ 50 |
|-------------|-----------------------|-----------------------------------|---------------------|--------------------------|-----------------------------|---|-----------------------------|--------|----------------------|
| Switzerland | 190,82 | 110,96 | 78,65 | 74,47 | 131,75 | 8,42 | 1 | 1 | 1 |
| Denmark | 190,01 | 94,73 | 73,28 | 79,96 | 91,67 | 6,66 | 1 | 1 | 1 |
| Netherlands | 183,31 | 83,89 | 72,78 | 75,76 | 78,64 | 7,35 | 1 | 0 | 1 |
| Finland | 182,79 | 89,05 | 72,99 | 76,4 | 77,46 | 8,64 | 1 | 1 | 1 |
| Austria | 182,37 | 78,23 | 74,77 | 78,4 | 75,49 | 10,4 | 1 | 1 | 1 |
| Australia | 181,52 | 99,29 | 57,56 | 77,71 | 84,14 | 7,38 | 1 | 0 | 1 |
| Iceland | 179,1 | 71,88 | 75,87 | 65,69 | 96,77 | 6,31 | 1 | 1 | 1 |
| Germany | 176,76 | 93,72 | 64,58 | 73,77 | 70,62 | 9,12 | 1 | 1 | 1 |

| | | | | | | | | | |
|----------------------|--------|--------|-------|-------|--------|-------|---|---|---|
| New Zealand | 175,77 | 81,44 | 57,74 | 73,58 | 79,14 | 8,09 | 1 | 0 | 1 |
| Norway | 173,57 | 79,43 | 66,65 | 75,5 | 106,09 | 8,49 | 1 | 1 | 1 |
| Estonia | 173,56 | 61,22 | 76,62 | 72,83 | 56,45 | 9,11 | 1 | 1 | 1 |
| Oman | 172,08 | 84,45 | 79,74 | 58,42 | 50,18 | 5,75 | 1 | 1 | 1 |
| Sweden | 171,4 | 90,55 | 52,8 | 68,8 | 79,17 | 8,56 | 1 | 1 | 1 |
| Slovenia | 168,2 | 56,14 | 78,21 | 65,28 | 59,38 | 10,89 | 1 | 1 | 1 |
| United States | 166,98 | 102,58 | 52,26 | 69,03 | 71,92 | 3,99 | 1 | 1 | 1 |
| Spain | 164,48 | 62,68 | 66,87 | 78,8 | 59,09 | 9,59 | 1 | 0 | 1 |
| Japan | 162,32 | 76,53 | 78,05 | 80,68 | 87,77 | 13,04 | 1 | 1 | 1 |
| Portugal | 161,91 | 44,96 | 70,11 | 71,93 | 52,88 | 12,65 | 1 | 0 | 1 |
| Lithuania | 160,02 | 54,6 | 66,16 | 70,97 | 47,66 | 10,97 | 1 | 1 | 1 |
| Canada | 159,99 | 82,76 | 58,81 | 71,8 | 70,08 | 7,53 | 1 | 1 | 1 |
| United Kingdom | 158,99 | 82,56 | 54,74 | 74,93 | 71,03 | 9,57 | 1 | 0 | 1 |
| Czech Republic | 156,33 | 56,4 | 74,69 | 75,37 | 49,18 | 15,46 | 1 | 1 | 1 |
| Croatia | 156,1 | 43,13 | 75,09 | 64,23 | 56,36 | 14,2 | 1 | 0 | 1 |
| United Arab Emirates | 156,03 | 85,74 | 84,65 | 68,02 | 61,67 | 4,93 | 0 | 1 | 1 |
| Qatar | 154,58 | 93,12 | 87,71 | 73 | 64,23 | 5,97 | 0 | 1 | 1 |
| Belgium | 150,89 | 76,54 | 55,83 | 75,2 | 78,52 | 6,91 | 0 | 1 | 1 |
| Ireland | 150,89 | 73,84 | 54,98 | 52,78 | 83,11 | 7,4 | 1 | 0 | 1 |
| France | 150,73 | 74,06 | 50,8 | 80,99 | 80,62 | 13,48 | 1 | 0 | 1 |
| Slovakia | 149,68 | 50,39 | 70,26 | 60,89 | 49,08 | 11,14 | 1 | 1 | 1 |

| | | | | | | | | | |
|-------------------------|--------|-------|-------|-------|-------|-------|---|---|---|
| Latvia | 147,59 | 45,94 | 62,79 | 62,18 | 52,87 | 8,48 | 1 | 1 | 1 |
| Saudi Arabia | 147,37 | 88,18 | 74,07 | 60,71 | 49,95 | 2,57 | 0 | 1 | 1 |
| Israel | 144,14 | 70,36 | 68,79 | 73,76 | 86,63 | 13,68 | 0 | 0 | 1 |
| Cyprus | 144,06 | 52,34 | 68,03 | 52,71 | 64,05 | 7,12 | 0 | 0 | 1 |
| Singapore | 143,82 | 83,62 | 67,02 | 70,88 | 85,59 | 19,38 | 1 | 1 | 1 |
| Taiwan | 138,82 | 56,08 | 84,76 | 86,39 | 65,25 | 23,63 | 0 | 0 | 1 |
| Italy | 138,63 | 58,7 | 55,63 | 66,77 | 73,11 | 9,52 | 0 | 1 | 1 |
| Hungary | 134,01 | 49,81 | 65,19 | 51,64 | 42,75 | 11,6 | 1 | 0 | 1 |
| Poland | 132,65 | 47,09 | 70,68 | 58,25 | 42,04 | 14,03 | 0 | 0 | 1 |
| Bulgaria | 126,34 | 40,26 | 61,88 | 56,13 | 40,92 | 8,97 | 0 | 1 | 2 |
| Turkey | 125,97 | 35,35 | 60,47 | 70,71 | 31,48 | 8,21 | 0 | 0 | 2 |
| Bosnia And Herzegovi na | 125,71 | 42,56 | 56,99 | 53,48 | 38,56 | 10,75 | 0 | 1 | 2 |
| Uruguay | 124,63 | 30,62 | 48,78 | 67,66 | 51,09 | 17,33 | 1 | 0 | 2 |
| Costa Rica | 121,65 | 39,01 | 44,66 | 62,92 | 50,64 | 10,07 | 1 | 0 | 2 |
| Mexico | 120,29 | 35,54 | 45 | 72,51 | 35,05 | 10,46 | 0 | 0 | 2 |
| Georgia | 118,87 | 25,67 | 77,38 | 53,83 | 28,05 | 14,16 | 0 | 0 | 2 |
| Ecuador | 118,76 | 30,54 | 45,59 | 68,81 | 40,08 | 12,59 | 0 | 0 | 2 |
| Malaysia | 116,94 | 56,1 | 42,11 | 69,59 | 39,51 | 9,92 | 0 | 1 | 2 |
| Serbia | 116,24 | 35,83 | 61,74 | 51,62 | 39,19 | 17,08 | 0 | 1 | 2 |
| Jordan | 114,04 | 33,13 | 59,27 | 65,38 | 52,17 | 8,03 | 0 | 0 | 2 |
| Kuwait | 113,98 | 75,32 | 65,96 | 58,97 | 49,4 | 13,18 | 0 | 1 | 2 |
| Argentina | 110,5 | 34,56 | 36,69 | 68,58 | 37,87 | 22,68 | 0 | 0 | 2 |
| Panama | 110,32 | 32,05 | 54,15 | 60,43 | 54,07 | 12,33 | 0 | 1 | 2 |

| | | | | | | | | | |
|-----------------|--------|-------|-------|-------|-------|-------|---|---|---|
| North Macedonia | 108,47 | 31,97 | 61,41 | 56,49 | 34,15 | 12,67 | 0 | 1 | 2 |
| Ukraine | 107,35 | 30,1 | 51,72 | 53,43 | 29,21 | 10,77 | 0 | 1 | 2 |
| Belarus | 107,35 | 31,96 | 39,73 | 44,44 | 30,69 | 16,79 | 1 | 1 | 2 |
| Pakistan | 105,14 | 27,28 | 56,14 | 60,52 | 21,59 | 12,62 | 0 | 1 | 2 |
| Brazil | 104,75 | 28,66 | 32,15 | 57,33 | 31,82 | 16,13 | 0 | 0 | 2 |
| India | 104,52 | 47,13 | 55,28 | 66,25 | 25,16 | 11,31 | 0 | 1 | 2 |
| Morocco | 104,42 | 32,17 | 50,9 | 45,81 | 36,85 | 13,49 | 0 | 0 | 2 |
| China | 103,15 | 57,47 | 69,83 | 66,38 | 42,54 | 29,02 | 0 | 1 | 2 |
| Azerbaijan | 102,88 | 25,77 | 68,34 | 44,02 | 30,9 | 16,58 | 0 | 0 | 2 |
| Russia | 101,67 | 34,61 | 59,87 | 58,44 | 33,17 | 11,14 | 0 | 1 | 2 |
| Colombia | 101,33 | 26,13 | 44,23 | 66,72 | 31 | 18,67 | 0 | 0 | 2 |
| Thailand | 100,29 | 31,36 | 60,62 | 78,08 | 49,32 | 22,16 | 0 | 1 | 2 |
| Chile | 99,9 | 33,33 | 47,92 | 63,72 | 49,33 | 17,6 | 0 | 0 | 2 |
| Hong Kong | 99,34 | 62,45 | 78,27 | 66,34 | 79,94 | 45,19 | 0 | 1 | 2 |
| Kazakhstan | 96,42 | 34,92 | 45,98 | 60,09 | 29,76 | 9,63 | 0 | 1 | 2 |
| Lebanon | 94,47 | 29,49 | 53,04 | 63,29 | 75,88 | 14,07 | 0 | 0 | 2 |
| Indonesia | 90,1 | 20,89 | 53,77 | 60,49 | 37,44 | 22,01 | 0 | 1 | 2 |
| Vietnam | 88,38 | 27,26 | 53,91 | 58,28 | 38,05 | 20,59 | 0 | 1 | 2 |
| Egypt | 86,31 | 19,99 | 53,21 | 46,21 | 30,1 | 13,65 | 0 | 0 | 2 |
| Peru | 83,3 | 28,81 | 33,39 | 56,38 | 36,26 | 16,88 | 0 | 0 | 2 |
| Sri Lanka | 79,78 | 20,77 | 58,81 | 72,63 | 31,28 | 35,01 | 0 | 1 | 2 |
| Philippines | 78,39 | 19,71 | 57,84 | 67,09 | 40,65 | 30,14 | 0 | 1 | 2 |
| Kenya | 75,77 | 27,67 | 38,60 | 55,83 | 36,3 | 27,27 | 0 | 0 | 2 |

| | | | | | | | | | |
|-------------|--------|-------|-------|-------|-------|-------|---|---|---|
| Bangladesh | 65,27 | 24,07 | 35,78 | 42,7 | 33,31 | 14,43 | 0 | 1 | 2 |
| Iran | 64,73 | 19,00 | 50,81 | 52,25 | 41,34 | 34,24 | 0 | 1 | 2 |
| Nigeria | 52 | 9,78 | 36,14 | 48,89 | 29,74 | 17,74 | 0 | 1 | 2 |
| Puerto Rico | 130,77 | 69,22 | 36,65 | 55,57 | 69,51 | 3,56 | 0 | 1 | |
| South Korea | 130,02 | 76,6 | 73,14 | 82,34 | 81,2 | 23,63 | 0 | 1 | |

2. Дискриминантный анализ

2.1. Выделить 1-3 наблюдения, подлежащих дискриминации

Наблюдения, которые будут подлежать дискриминации - данные о Южной Корее и Пуэрто-Рико, так как данные страны находились на стыке первой и второй половина топа 100 лучших стран мира для проживания, а значит они наиболее удалены от центра двух выделенных кластеров и без анализа их сложно отнести к одной из исследуемых групп.

2.2. Провести дискриминантный анализ

Дискриминантный анализ проводился с помощью компьютерной программы для статистической обработки данных - SPSS.

2.3. Выражение для дискриминантной функции

Таблица 2. Коэффициенты канонической дискриминантной функции

| | Функция 1 |
|---|--------------|
| Индекс качества жизни | ,029 |
| Паритет покупательной способности | ,018 |
| Индекс безопасности | ,033 |
| Индекс здравоохранения | -,008 |
| Индекс стоимости проживания | ,011 |
| Соотношения стоимости жилья и заработка | -,012 |
| Степень загрязнения воздуха (0 - серьезное, 1 - не серьезное) | ,837 |
| Климат (1- благоприятный климат, 0 - не благоприятный) | -,235 |
| (Константа) | -6,907 |
| Нестандартизованные коэффициенты | |

Дискриминантная функция: $f(x) = -6,907 + 0,29x_1 + 0,18x_2 + 0,33x_3 - 0,008x_4 + 0,011x_5 - 0,012x_6 + 0,837x_7 - 0,235x_8$

Где $x_1...x_8$ - это все независимые переменные от индекса качества жизни до климата.

2.4. Оценка значимости дискриминантной функции

Таблица 3. Лямбда Уилкса

| Критерий для функций | Лямбда Уилкса | Хи-квадрат | ст.св. | знач. |
|----------------------|---------------|------------|--------|-------|
| 1 | ,203 | 113,333 | 8 | ,000 |

В данном примере лямбда Уилкса равна 0,203, что является хорошим результатом, однако было бы лучше, если бы она была ближе к 0.

Функция значима, так как уровень значимости меньше 0,05.

2.5. Определение относительного вклада каждой переменной в формирование классов

Таблица 4. Коэффициенты стандартизированной канонической дискриминантной функции

| Функция 1 | |
|---|-------|
| Индекс качества жизни | ,492 |
| Паритет покупательной способности | ,274 |
| Индекс безопасности | ,353 |
| Индекс здравоохранения | -,065 |
| Индекс стоимости проживания | ,169 |
| Соотношения стоимости жилья и заработка | -,081 |
| Степень загрязнения воздуха (0 - серьезное, 1 - не серьезное) | ,300 |
| Климат (1- благоприятный климат, 0 - не благоприятный) | -,117 |

Относительный вклад каждой переменной в различие двух сформированных классов можно оценить с помощью коэффициентов стандартизированной канонической дискриминантной функции. Чем больше коэффициент по модулю, тем больше влияние на формирование классов оказывает переменная.

Таким образом индекс качества жизни, индекс безопасности и степень загрязнения воздуха имеют наибольший вклад в формирование классов, а индекс здравоохранения и соотношение стоимости жилья и заработка наименьший.

2.6. Определение средних значений дискриминантной функции по группам

Таблица 5. Функции в центроидах групп

| Функция | |
|-----------|--------|
| Кластер 1 | 1 |
| 1 | 1,983 |
| 2 | -1,932 |

Нестандартизованные канонические дискриминантные функции, вычисленные в групповых средних

Расстояние между центроидами – среднее значение дискриминантной функции в исследуемых группах. В группе 1 - страны более комфортные для жизни, среднее значение дискриминантной функции равно = 1,983. В группе 2 - страны менее комфортные для жизни, среднее значение дискриминантной функции равно = -1,932.

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

Таблица 6. Статистика по наблюдениям

| | | Фактическая группа | Предсказанная группа | P(D>d G=g) | | P(G=g D=d) |
|----------|---|--------------------|----------------------|--------------|--------|--------------|
| | | | | PM | ст.св. | |
| Исходный | 1 | 1 | 1 | ,009 | 1 | 1,000 |
| | 2 | 1 | 1 | ,093 | 1 | 1,000 |

| | | | | | |
|----|---|-----|------|---|-------|
| 3 | 1 | 1 | ,163 | 1 | 1,000 |
| 4 | 1 | 1 | ,227 | 1 | 1,000 |
| 5 | 1 | 1 | ,314 | 1 | 1,000 |
| 6 | 1 | 1 | ,249 | 1 | 1,000 |
| 7 | 1 | 1 | ,226 | 1 | 1,000 |
| 8 | 1 | 1 | ,438 | 1 | 1,000 |
| 9 | 1 | 1 | ,520 | 1 | 1,000 |
| 10 | 1 | 1 | ,382 | 1 | 1,000 |
| 11 | 1 | 1 | ,716 | 1 | 1,000 |
| 12 | 1 | 1 | ,359 | 1 | 1,000 |
| 13 | 1 | 1 | ,759 | 1 | 1,000 |
| 14 | 1 | 1 | ,813 | 1 | 1,000 |
| 15 | 1 | 1 | ,727 | 1 | 1,000 |
| 16 | 1 | 1 | ,990 | 1 | 1,000 |
| 17 | 1 | 1 | ,562 | 1 | 1,000 |
| 18 | 1 | 1 | ,748 | 1 | ,998 |
| 19 | 1 | 1 | ,548 | 1 | ,995 |
| 20 | 1 | 1 | ,943 | 1 | ,999 |
| 21 | 1 | 1 | ,965 | 1 | ,999 |
| 22 | 1 | 1 | ,641 | 1 | ,997 |
| 23 | 1 | 1 | ,778 | 1 | ,999 |
| 24 | 1 | 1 | ,887 | 1 | ,999 |
| 25 | 1 | 1 | ,980 | 1 | 1,000 |
| 26 | 1 | 1 | ,190 | 1 | ,926 |
| 27 | 1 | 1 | ,918 | 1 | ,999 |
| 28 | 1 | 1 | ,577 | 1 | ,996 |
| 29 | 1 | 1 | ,452 | 1 | ,991 |
| 30 | 1 | 1 | ,282 | 1 | ,969 |
| 31 | 1 | 1 | ,456 | 1 | ,991 |
| 32 | 1 | 1 | ,348 | 1 | ,982 |
| 33 | 1 | 1 | ,199 | 1 | ,933 |
| 34 | 1 | 1 | ,818 | 1 | ,999 |
| 35 | 1 | 1 | ,207 | 1 | ,938 |
| 36 | 1 | 2** | ,058 | 1 | ,560 |
| 37 | 1 | 1 | ,247 | 1 | ,958 |
| 38 | 1 | 2** | ,054 | 1 | ,529 |
| 40 | 2 | 2 | ,294 | 1 | ,972 |
| 41 | 2 | 2 | ,320 | 1 | ,978 |
| 42 | 2 | 2 | ,135 | 1 | ,859 |
| 43 | 2 | 2 | ,123 | 1 | ,837 |
| 44 | 2 | 2 | ,712 | 1 | ,998 |
| 45 | 2 | 2 | ,211 | 1 | ,941 |
| 46 | 2 | 2 | ,755 | 1 | ,998 |

| | | | | | |
|----|-------------------------|---|------|---|-------|
| 47 | 2 | 2 | ,704 | 1 | ,998 |
| 48 | 2 | 2 | ,485 | 1 | ,993 |
| 49 | 2 | 2 | ,376 | 1 | ,985 |
| 50 | 2 | 2 | ,116 | 1 | ,819 |
| 51 | 2 | 2 | ,763 | 1 | 1,000 |
| 52 | 2 | 2 | ,721 | 1 | ,998 |
| 53 | 2 | 2 | ,723 | 1 | ,998 |
| 54 | 2 | 2 | ,969 | 1 | 1,000 |
| 55 | 2 | 2 | ,659 | 1 | ,997 |
| 56 | 2 | 2 | ,869 | 1 | 1,000 |
| 57 | 2 | 2 | ,532 | 1 | 1,000 |
| 58 | 2 | 2 | ,881 | 1 | ,999 |
| 59 | 2 | 2 | ,822 | 1 | ,999 |
| 60 | 2 | 2 | ,458 | 1 | ,991 |
| 61 | 2 | 2 | ,576 | 1 | ,996 |
| 62 | 2 | 2 | ,885 | 1 | ,999 |
| 63 | 2 | 2 | ,632 | 1 | 1,000 |
| 64 | 2 | 2 | ,969 | 1 | 1,000 |
| 65 | 2 | 2 | ,970 | 1 | 1,000 |
| 66 | 2 | 2 | ,230 | 1 | ,951 |
| 67 | 2 | 2 | ,620 | 1 | 1,000 |
| 68 | 2 | 2 | ,813 | 1 | ,999 |
| 69 | 2 | 2 | ,457 | 1 | 1,000 |
| 70 | 2 | 2 | ,525 | 1 | 1,000 |
| 71 | 2 | 2 | ,603 | 1 | 1,000 |
| 72 | 2 | 2 | ,245 | 1 | 1,000 |
| 73 | 2 | 2 | ,231 | 1 | 1,000 |
| 74 | 2 | 2 | ,277 | 1 | 1,000 |
| 75 | 2 | 2 | ,176 | 1 | 1,000 |
| 76 | 2 | 2 | ,068 | 1 | 1,000 |
| 77 | 2 | 2 | ,096 | 1 | 1,000 |
| 78 | 2 | 2 | ,010 | 1 | 1,000 |
| 79 | не сгруппирован о | 2 | ,180 | 1 | ,918 |
| 80 | не сгруппирован о | 1 | ,115 | 1 | ,816 |

**. Ошибочно классифицированное наблюдение

У большинства классифицируемых объектов фактическая и предсказанные группы совпали. Исключением стали наблюдения 36 и 38. Не сгруппированное наблюдения 79 попало в группу 2, а объект 80 в группу 1.

Вероятности, с которыми наблюдения входят в предсказанные группы можно увидеть в столбце с названием Р.

2.8. Проверка значимости различий средних значений дискриминантной функции в двух группах

Таблица 7. Критерии равенства групповых средних

| | Лямбда Уилкса | F | ст.св.1 | ст.св.2 | знач. |
|---|------------------|---------|---------|---------|-------|
| Индекс качества жизни | ,233 | 246,984 | 1 | 75 | ,000 |
| Паритет покупательной способности | ,367 | 129,388 | 1 | 75 | ,000 |
| Индекс безопасности | ,633 | 43,415 | 1 | 75 | ,000 |
| Индекс здравоохранения | ,731 | 27,574 | 1 | 75 | ,000 |
| Индекс стоимости проживания | ,497 | 76,043 | 1 | 75 | ,000 |
| Соотношения стоимости жилья и заработка | ,744 | 25,814 | 1 | 75 | ,000 |
| Степень загрязнения воздуха (0 - серьезное, 1 - не серьезное) | ,515 | 70,533 | 1 | 75 | ,000 |
| Климат (1- благоприятный климат, 0 - не благоприятный) | ,995 | ,357 | 1 | 75 | ,552 |

Переменную климат можно убрать из анализа, т.к. значимость ее различия с другими значениями их группы больше 0,05.

Остальные переменные имеют значимость меньше или равно 0,05 и следовательно средние двух групп значимо различаются, т.е. доказано наличие дискриминирующих особенностей этих переменных.

2.9. Оценить качество дискриминантного анализа

Таблица 8. Собственные значения

| Функция | Собственное значение | % дисперсии | Суммарный % | Каноническая корреляция |
|---------|----------------------|-------------|-------------|-------------------------|
| 1 | 3,934 ^а | 100,0 | 100,0 | ,893 |

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

Собственное значение - отношение межгрупповой дисперсии к внутригрупповой дисперсии выборочных значений дискриминантной функции. Чем больше собственное значение, тем лучше подобрана дискриминантная функция.

В исследуемом случае собственное значение равно 3,934, что является достаточно большим значением. А значит дискриминантная функция была подобрана хорошо.

Каноническая корреляция характеризует качество достоверности дискриминации. В данном случае достоверность дискриминации равно 0,893, что является хорошим показателем, а значит функция была подобрана хорошо.

2.10. Оценить целесообразность проведения дискриминантного анализа по Вашим данным.

Я считаю, что было целесообразно проводить анализ по исследуемым данным, т.к. большинство показателей являются значимыми и позволяют достоверно определить кластеры, однако если убрать показатель климата, то анализ был бы еще более точным.

3. Построение дерева классификации

3.1. Выбор зависимой переменной

Так как задачей исследования является определение комфортности жизни в различных странах, в качестве зависимой переменной будет взята кластеризация, получившаяся в итоге дискриминантного анализа. (предсказанная группа из пункта 2.7)

3.2. Построение деревьев с помощью метода CHAID с различными независимыми переменными

Рассмотрим схемы деревьев с различными независимыми признаками.

- Независимый признак - индекс качества жизни

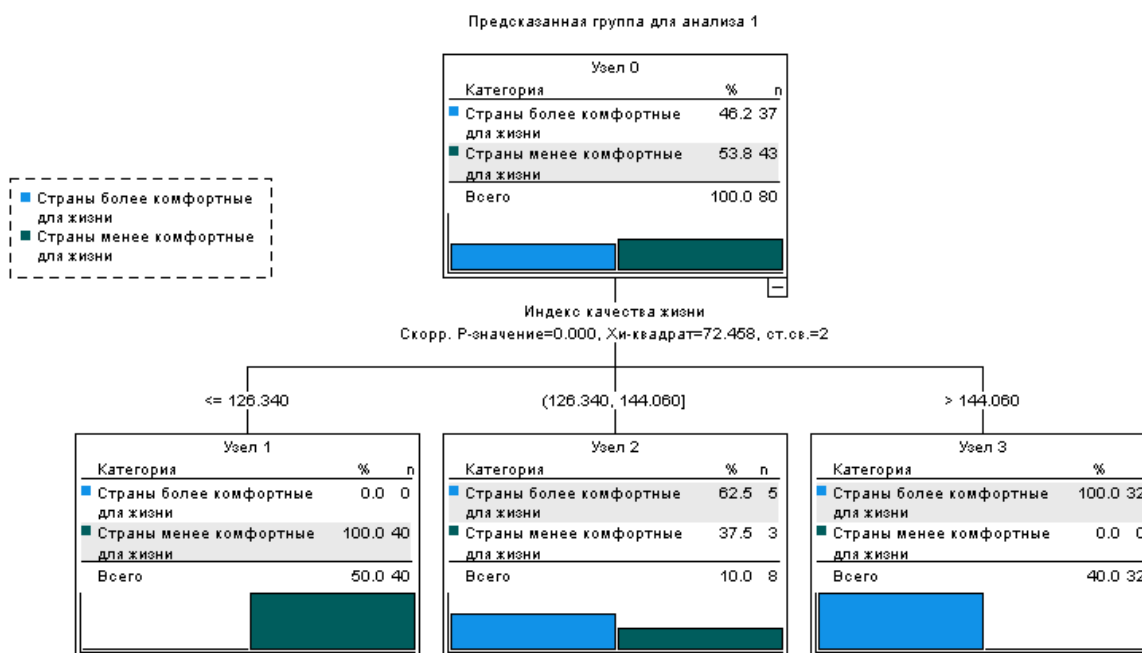


Рис.1. Дерево классификаций с независимой переменной - индекс качества жизни

- Независимый признак - индекс безопасности

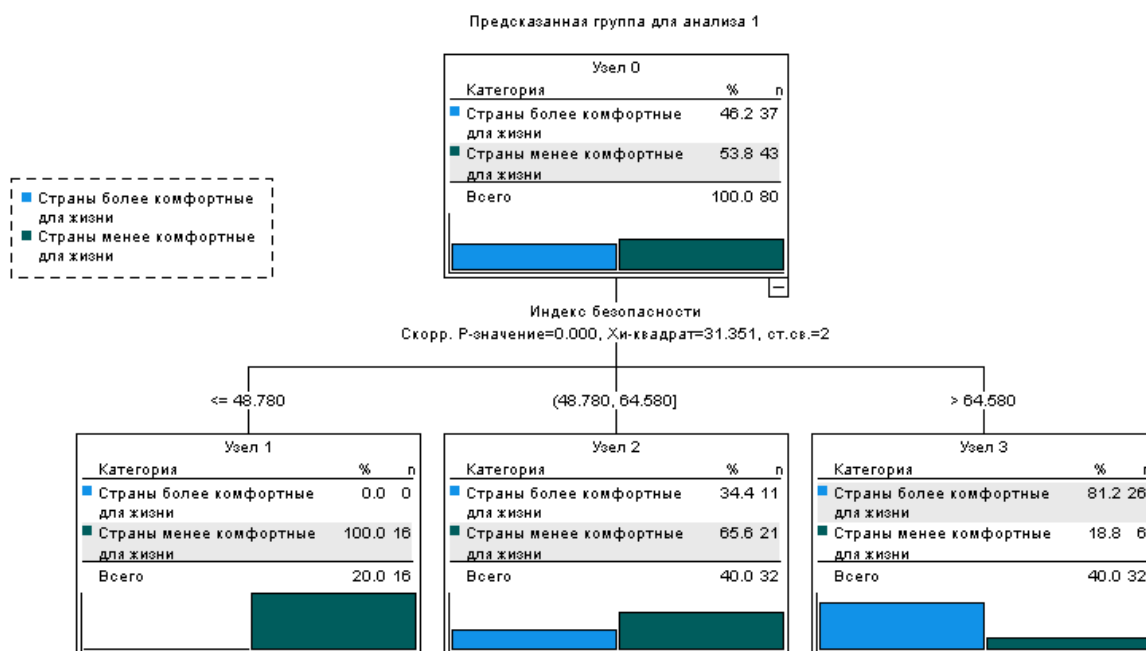


Рис.2. Дерево классификаций с независимой переменной - индекс безопасности

- Независимый признак - Степень загрязнения воздуха

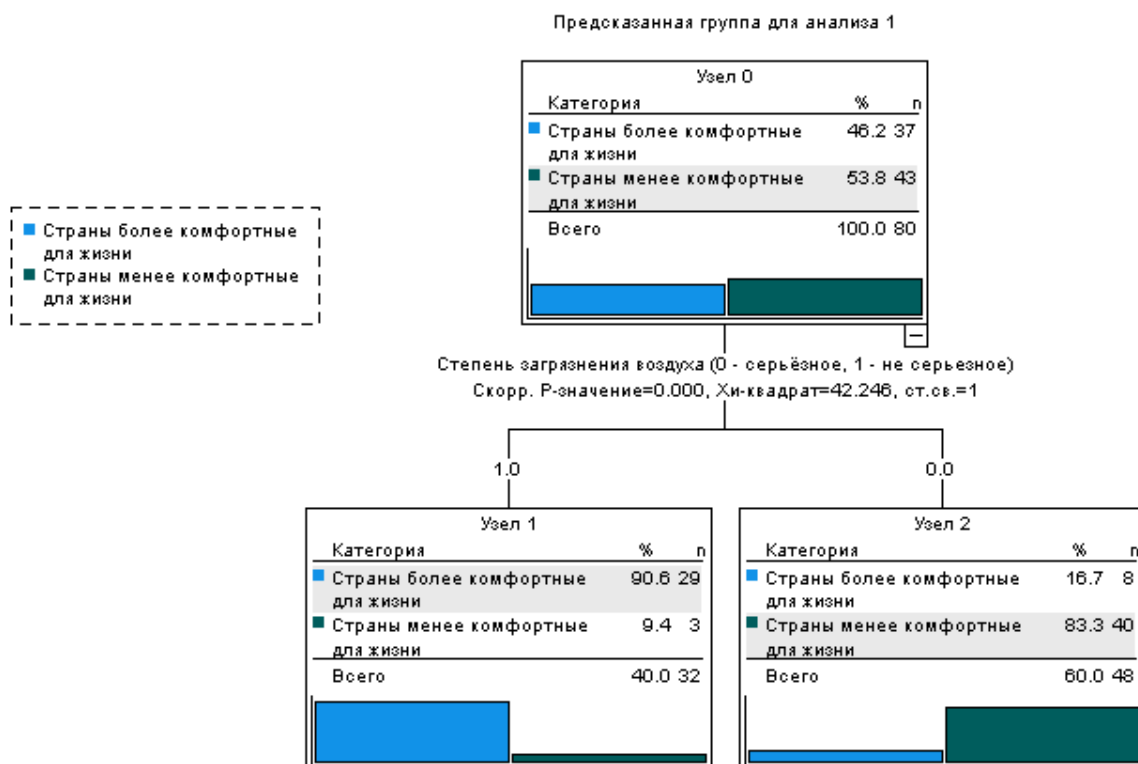



Рис.3. Дерево классификаций с независимой переменной - степень загрязнения воздуха

- Предсказанная группа для анализа 1

■ Страны более комфортные для жизни

■ Страны менее комфортные для жизни

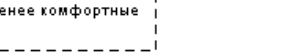
| Узел 0 | | |
|-----------------------------------|-------|----|
| Категория | % | n |
| Страны более комфортные для жизни | 46.2 | 37 |
| Страны менее комфортные для жизни | 53.8 | 43 |
| Всего | 100.0 | 80 |



Паритет покупательной способности
Скорр. Р-значение=0.000, Хи-квадрат=52.680, ст.св.=2


≤ 35.830000000000000000

| Узел 1 | | |
|-----------------------------------|-------|----|
| Категория | % | n |
| Страны более комфортные для жизни | 0.0 | 0 |
| Страны менее комфортные для жизни | 100.0 | 32 |
| Всего | 40.0 | 32 |




(35.830000000000000000, 71.8800000000000000]

| Узел 2 | | |
|-----------------------------------|------|----|
| Категория | % | n |
| Страны более комфортные для жизни | 58.3 | 14 |
| Страны менее комфортные для жизни | 41.7 | 10 |
| Всего | 30.0 | 24 |

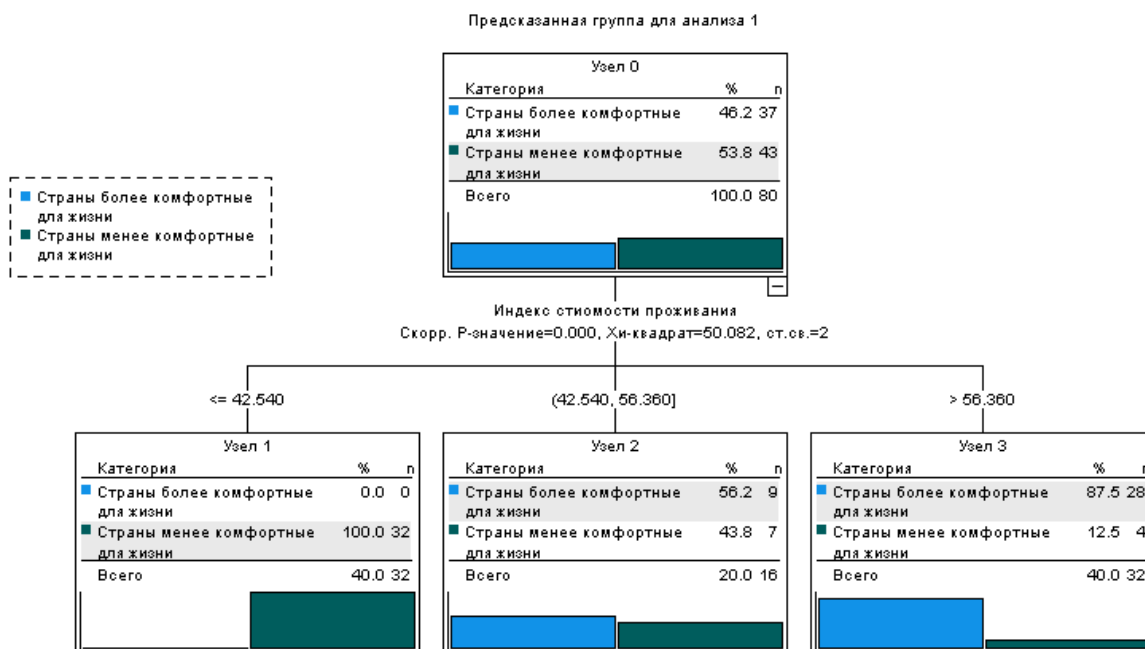


> 71.880000000000000000

| Узел 3 | | |
|-----------------------------------|------|----|
| Категория | % | n |
| Страны более комфортные для жизни | 95.8 | 23 |
| Страны менее комфортные для жизни | 4.2 | 1 |
| Всего | 30.0 | 24 |



- Независимый признак - индекс стоимости проживания



17

- Независимый признак - климат

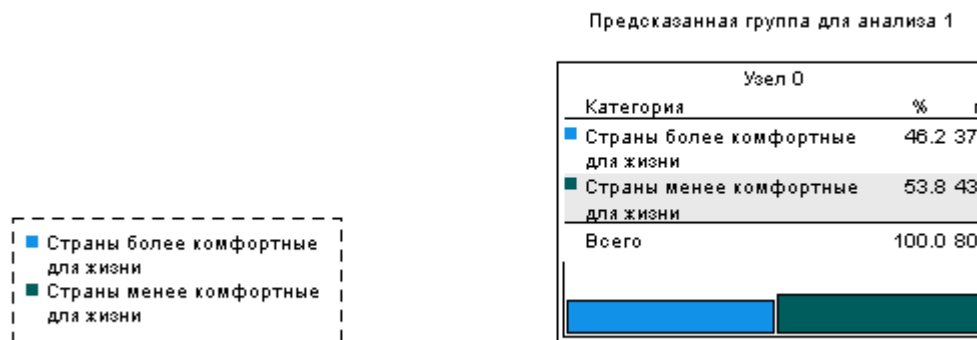


Рис.6. Дерево классификаций с независимой переменной - климат

- Независимый признак - соотношение стоимости жилья и заработка

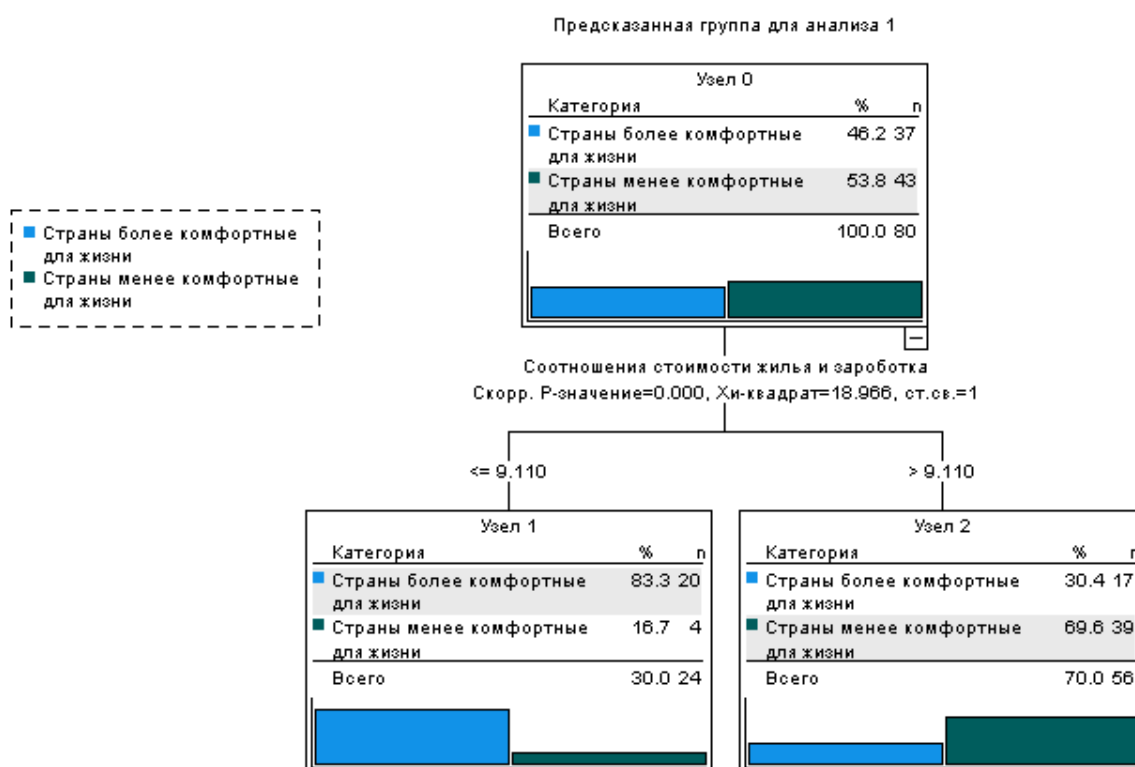


Рис.7. Дерево классификаций с независимой переменной - соотношение стоимости жилья и заработной платы

- Независимый признак - индекс здравоохранения

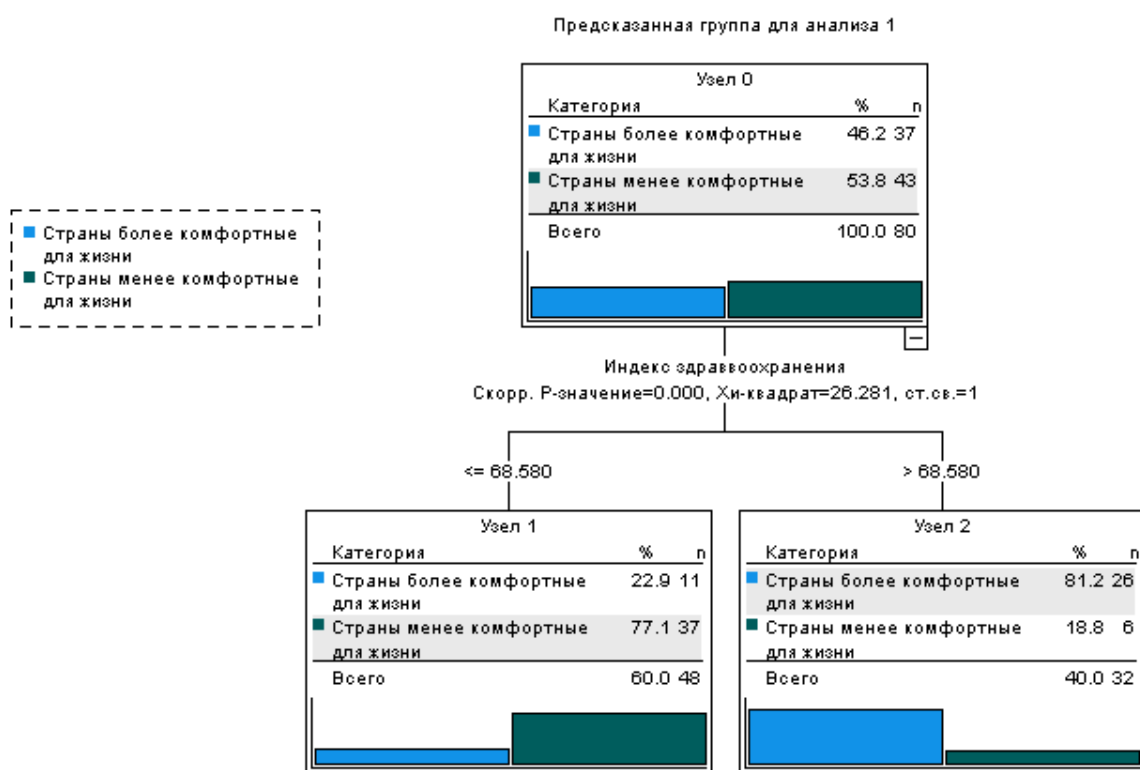


Рис.8. Дерево классификаций с независимой переменной - индекс здравоохранения

3.3. Выбор оптимального дерева с помощью таблицы классификации

- Независимый признак - индекс качества жизни

Таблица 9. Классификация с независимой переменной - индекс качества жизни

| Наблюдаемые | Предсказанные | | |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | Процент правильных |
| Страны более комфортные для жизни | 37 | 0 | 100,0% |
| Страны менее комфортные для жизни | 3 | 40 | 93,0% |
| Общая процентная доля | 50,0% | 50,0% | 96,3% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - индекс безопасности

Таблица 10. Классификация с независимой переменной - индекс безопасности

| Наблюдаемые | Предсказанные | | |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | Процент правильных |
| Страны более комфортные для жизни | 26 | 11 | 70,3% |
| Страны менее комфортные для жизни | 6 | 37 | 86,0% |
| Общая процентная доля | 40,0% | 60,0% | 78,8% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - Степень загрязнения воздуха

Таблица 11. Классификация с независимой переменной - степень загрязнения воздуха

Классификация

| Наблюдаемые | Предсказанные | | |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | Процент правильных |
| Страны более комфортные для жизни | 29 | 8 | 78,4% |
| Страны менее комфортные для жизни | 3 | 40 | 93,0% |
| Общая процентная доля | 40,0% | 60,0% | 86,3% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - Паритет покупательской способности

Таблица 12. Классификация с независимой переменной - паритет покупательской способности

| Наблюдаемые | Предсказанные | | |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | Процент правильных |
| Страны более комфортные для жизни | 37 | 0 | 100,0% |
| Страны менее комфортные для жизни | 11 | 32 | 74,4% |
| Общая процентная доля | 60,0% | 40,0% | 86,3% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - индекс стоимости проживания

Таблица 13. Классификация с независимой переменной - индекс стоимости проживания

| Наблюдаемые | Предсказанные | | Процент правильных |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | |
| Страны более комфортные для жизни | 37 | 0 | 100,0% |
| Страны менее комфортные для жизни | 11 | 32 | 74,4% |
| Общая процентная доля | 60,0% | 40,0% | 86,3% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - климат

Таблица 14. Классификация с независимой переменной - климат

| Наблюдаемые | Предсказанные | | Процент правильных |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | |
| Страны более комфортные для жизни | 0 | 37 | 0,0% |
| Страны менее комфортные для жизни | 0 | 43 | 100,0% |
| Общая процентная доля | 0,0% | 100,0% | 53,8% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - соотношение стоимости жилья и заработка

Таблица 15. Классификация с независимой переменной - соотношение стоимости жилья и заработка

| Наблюдаемые | Предсказанные | | Процент правильных |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | |
| Страны более комфортные для жизни | 20 | 17 | 54,1% |
| Страны менее комфортные для жизни | 4 | 39 | 90,7% |
| Общая процентная доля | 30,0% | 70,0% | 73,8% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

- Независимый признак - индекс здравоохранения

Таблица 16. Классификация с независимой переменной - индекс здравоохранения

| Наблюдаемые | Предсказанные | | Процент правильных |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | |
| Страны более комфортные для жизни | 26 | 11 | 70,3% |
| Страны менее комфортные для жизни | 6 | 37 | 86,0% |
| Общая процентная доля | 40,0% | 60,0% | 78,8% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

Вывод: независимые признаки индекс качества жизни, индекс стоимости проживания, паритет покупательской способности и степень загрязнения воздуха показали наилучшие результаты и процент правильных ответов в классификации по данным признакам превышает 85%, поэтому они будут использоваться для построения финального дерева решений.

3.4. Визуализация результатов

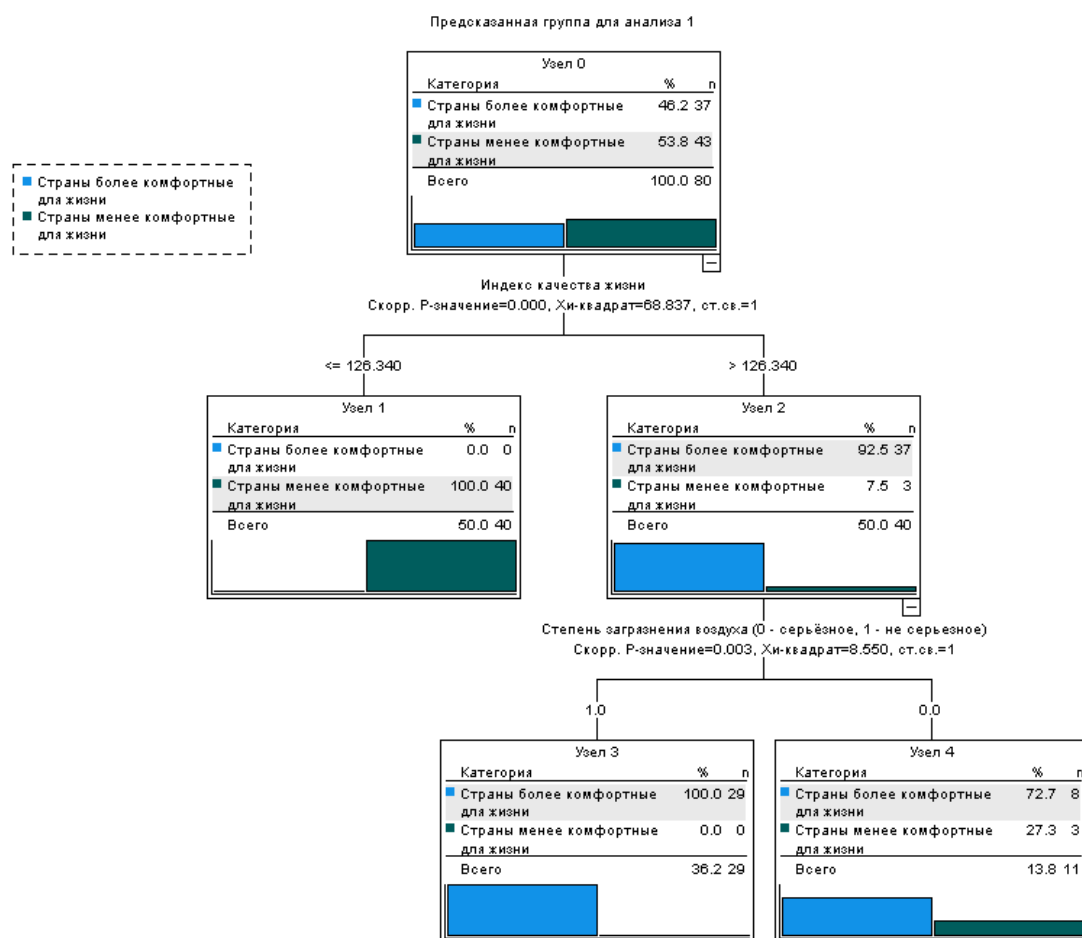


Рис.9. Дерево классификаций с независимыми переменными - индекс качества жизни, индекс стоимости проживания, паритет покупательской способности и степень загрязнения воздуха

Таблица 17. Классификация с независимыми переменными - индекс качества жизни, индекс стоимости проживания, паритет покупательской способности и степень загрязнения воздуха

| Наблюдаемые | Предсказанные | | |
|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|
| | Страны более комфортные для жизни | Страны менее комфортные для жизни | Процент правильных |
| Страны более комфортные для жизни | 37 | 0 | 100,0% |
| Страны менее комфортные для жизни | 3 | 40 | 93,0% |
| Общая процентная доля | 50,0% | 50,0% | 96,3% |

Метод построения: CHAID

Зависимая переменная: Предсказанная группа для анализа 1

3.5. Интерпретация результатов

Построение финального дерева показало, что достаточно использовать независимый признак индекса качества жизни, чтобы с максимально возможной вероятностью предсказать классификацию объекта. Однако, признаки индекс стоимости проживания, паритет покупательской способности и степень загрязнения воздуха могут также показать достаточно точный результат. Все остальные признаки не целесообразно использовать для классификации, так как вероятность верности их предсказания ниже 85%.

4. Декомпозиция смеси

Прологарифмируем исходные данные и построим эмпирическую гистограмму для более удобного анализа распределения.

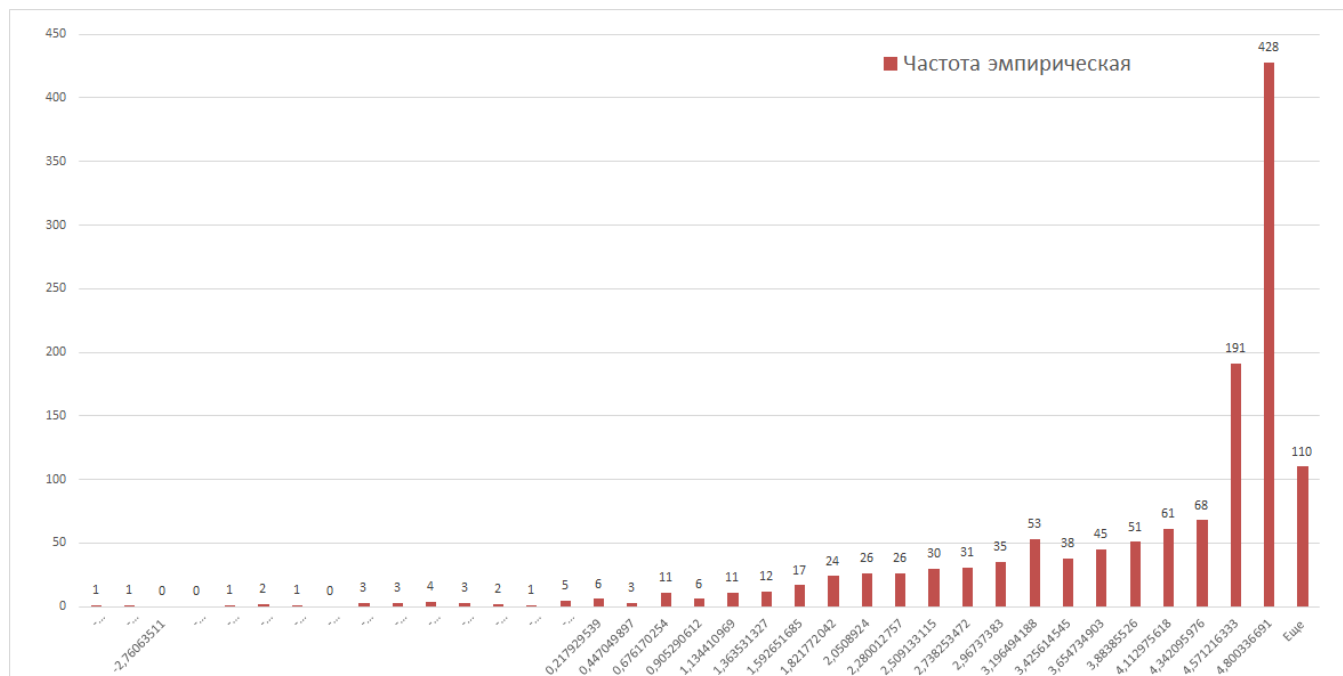


Рис.10. Гистограмма эмпирических частот прологарифмированных данных

На гистограмме можно выделить три части, предположительно являющиеся 3 разными смесями распределений. Проанализируем их по отдельности. Кластеризуем распределение на три составляющих: с начала до второй частоты 6 - первая часть; со второй частоты 11 до 68 - вторая; 191, 428, 110 - третья. (по частотам)

Выбираем "пики" по частотам из каждой составляющей распределения и записываем их. Находим где предположительно пересекаются хвосты распределений и вычитаем из этого значения мат.ожидание (ранее найденные "пики"). Предполагаем долю смеси в общем распределении. Ищем логнормальное распределение для нахождения теоретических частот до оптимизации. Строим гистограмму теоретических частот до оптимизации.

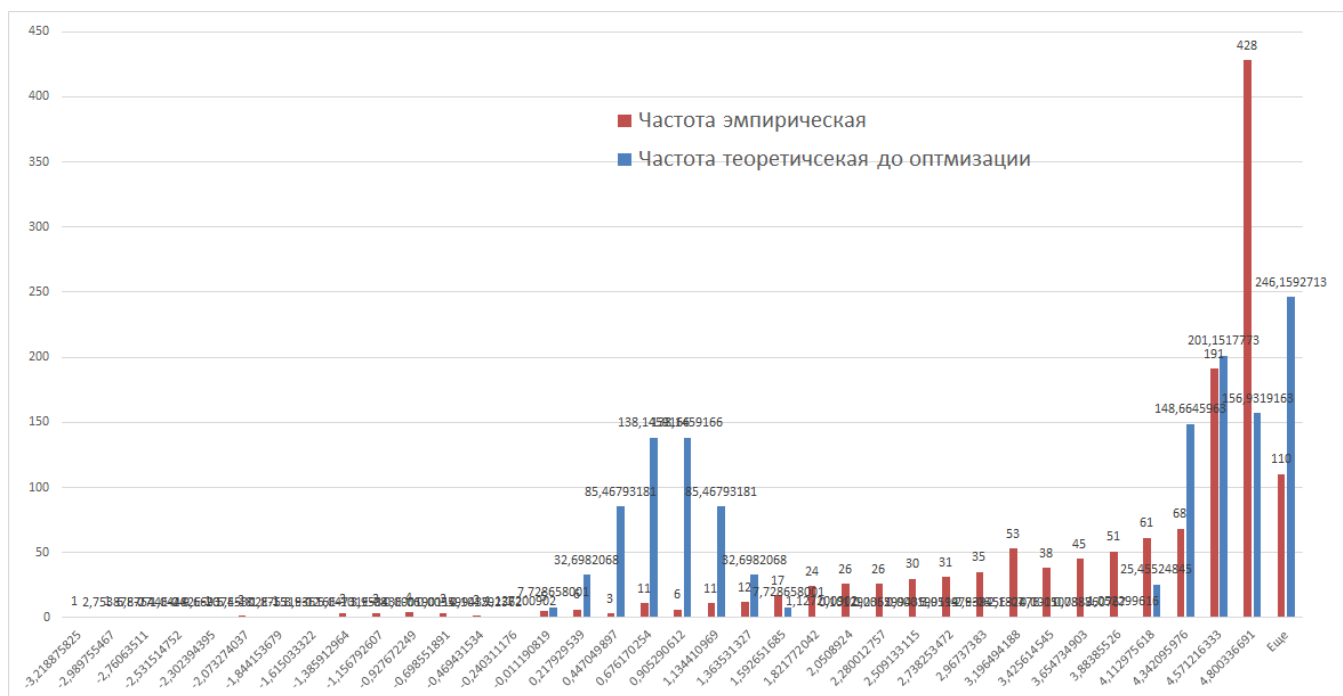


Рис.11. Гистограмма эмпирических и теоретических частот до оптимизации

На данном этапе можно заметить различия в гистограммах, которые показывают некоторые ошибочные теоретические предположения, особенно в зоне -0,11 - 1,36. Также значения в зоне 4,8 и больше оказались слишком малы. Оптимизируем теоретические частоты.

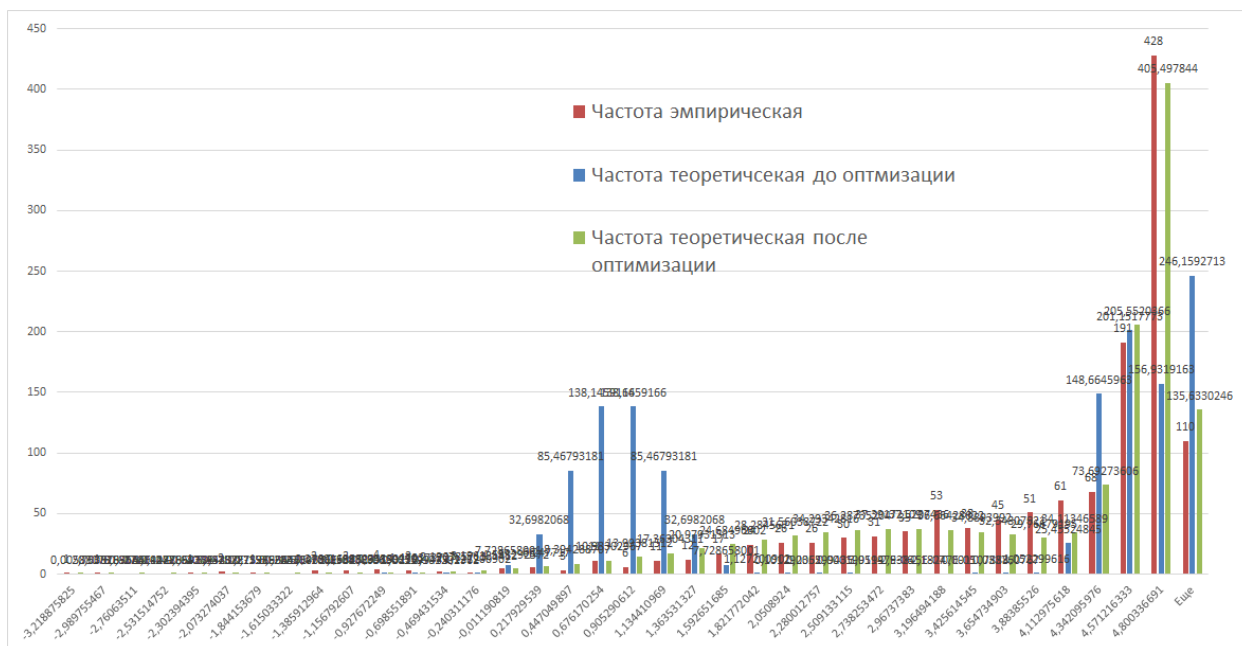


Рис.12. Гистограмма эмпирических частот, теоретических частот до оптимизации и после оптимизации

После оптимизации частот, значения получились максимально приближенными к искомым и на 90,95% совпали, а значит можно использовать найденные мат.ожидания, процентные соотношения и стандартные отклонения для исследования смесей трех распределений.

Найдем значения смесей на интервале $[-3; 5,5]$, а также значения функций принадлежности к стратам. Построим разные графики, для анализа смесей.

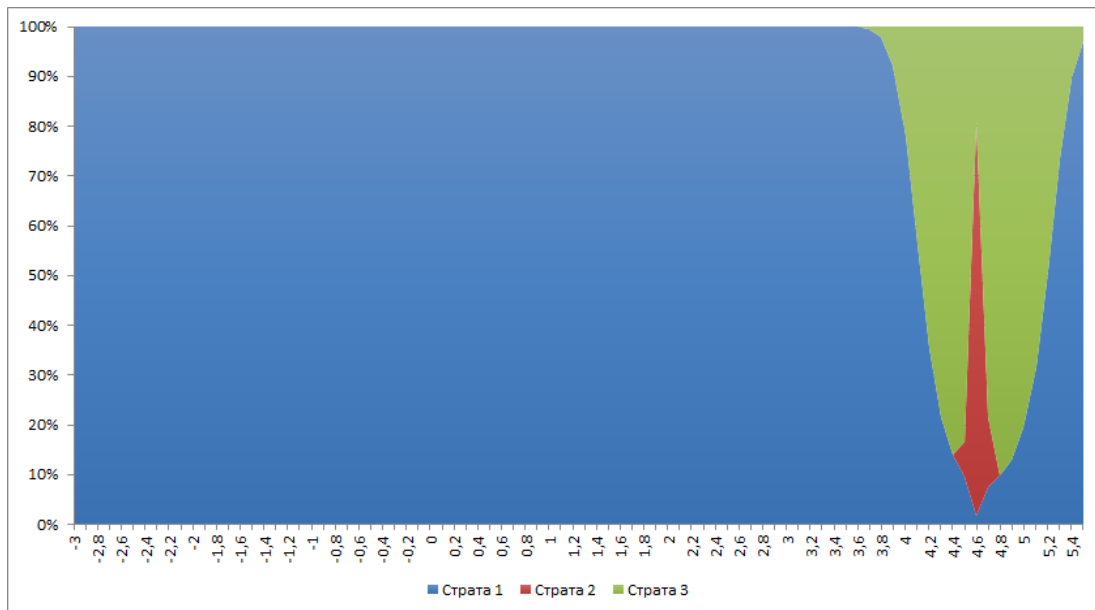


Рис.14. Гистограмма страт нормированная с областями и накоплениями

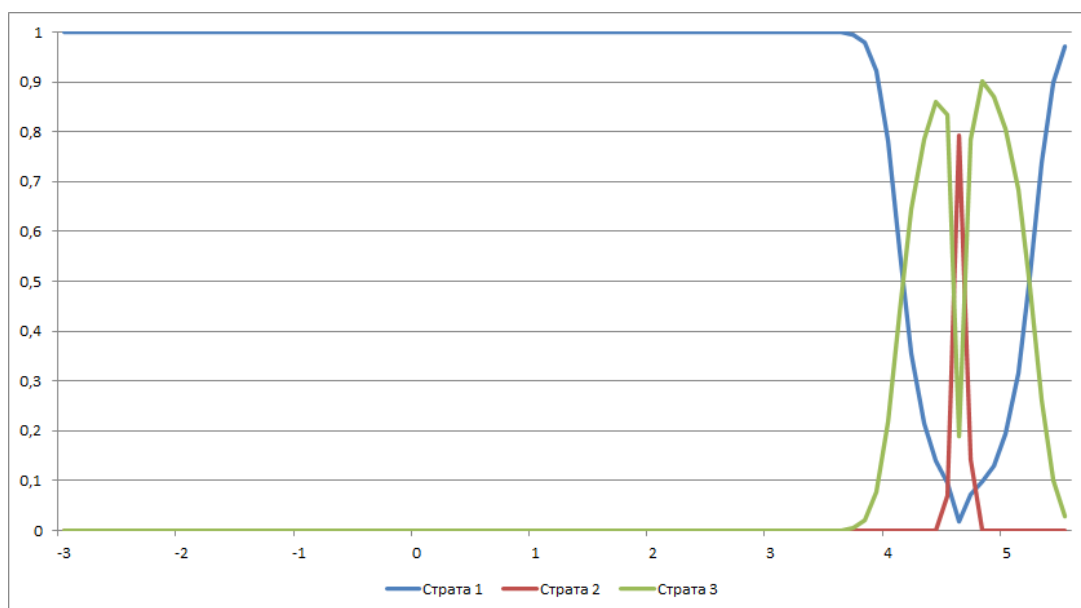


Рис.15. График страт

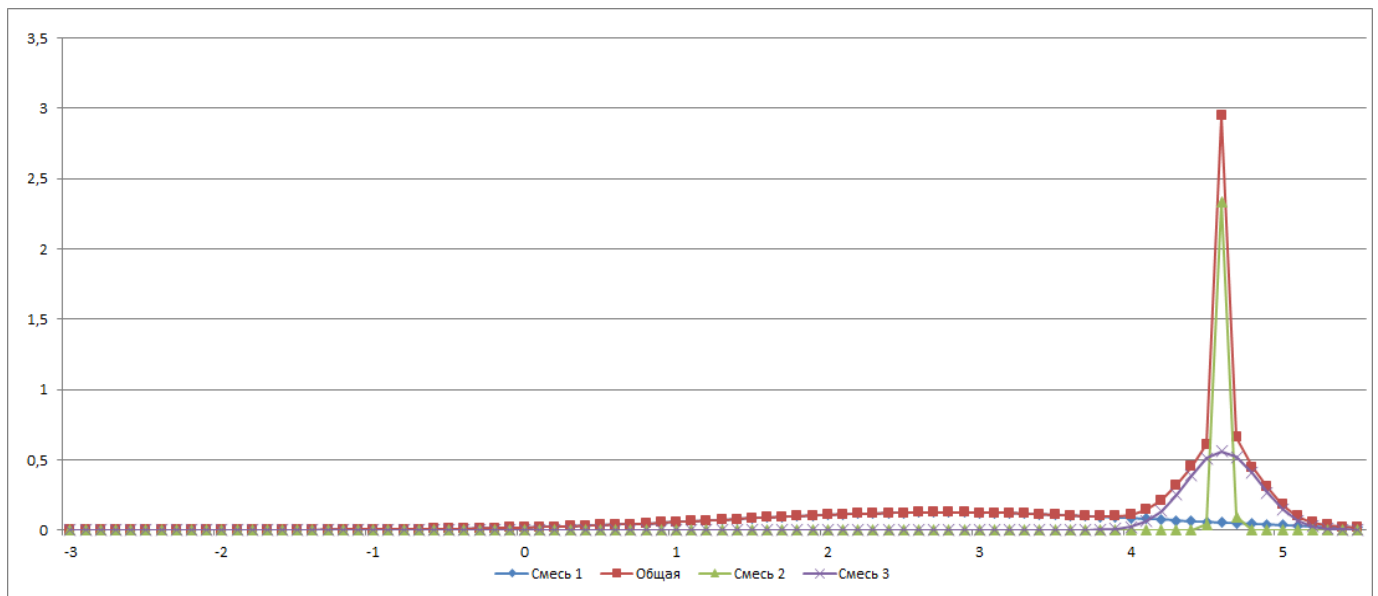


Рис.15. График с маркерами смесей распределений

Вывод: судя по графикам логарифмированные значения от -3 до 3,8 скорее всего попадут в группу 1. Значения от 3,8 до 4,4 будут принадлежать либо распределению 1, либо распределению 3. Значения от 4,4 до 4,8 с высокой вероятностью будут принадлежать либо распределению 2, либо распределению 3, хотя маленькая вероятность того, что они будут лежать в распределении 1, все же имеется. Значение от 4,8 до 5,5 имеют примерно одинаковую вероятность быть частью как и распределения 1, так и распределения 3.

Мой номер в группе - 8, значение, которое соответствует ему - 93,76. Логарифм этого числа равен 4,5407, значит скорее всего оно относится к распределению 3, однако есть небольшая вероятность, что это распределение 1 или 2, т.к. значение находится вблизи пересечения распределений.

Источники

1. https://www.numbeo.com/quality-of-life/rankings_by_country.jsp
2. <https://nafi.ru/upload/spss/NAFI%20-%202011.lection.pdf>