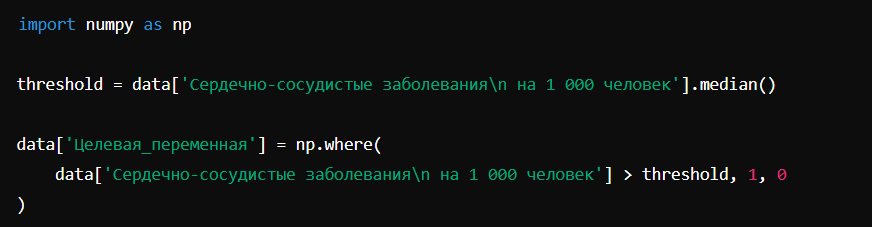
**3 Дерево решений**

**3.1 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:

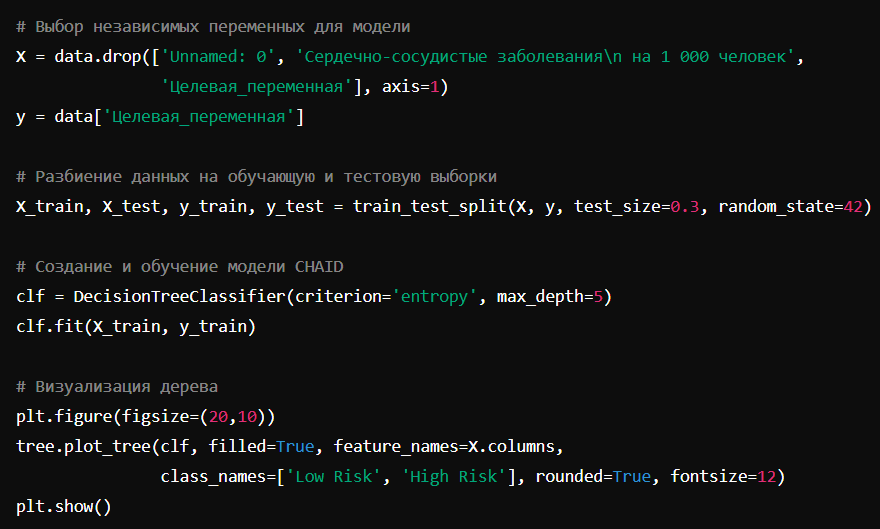
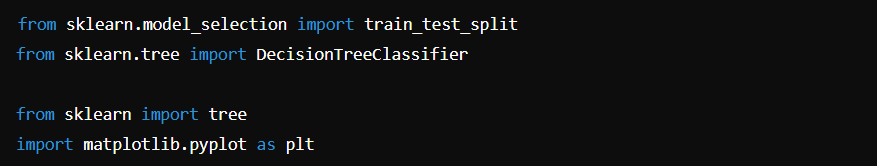
Для метода CHAID требуется дискретная зависимая переменная, поэтому мы преобразовали количественную переменную, указывающую количество сердечно-сосудистых заболеваний на 1000 человек, в бинарный формат. Это преобразование было выполнено с использованием медианы в качестве порога, что позволило разделить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

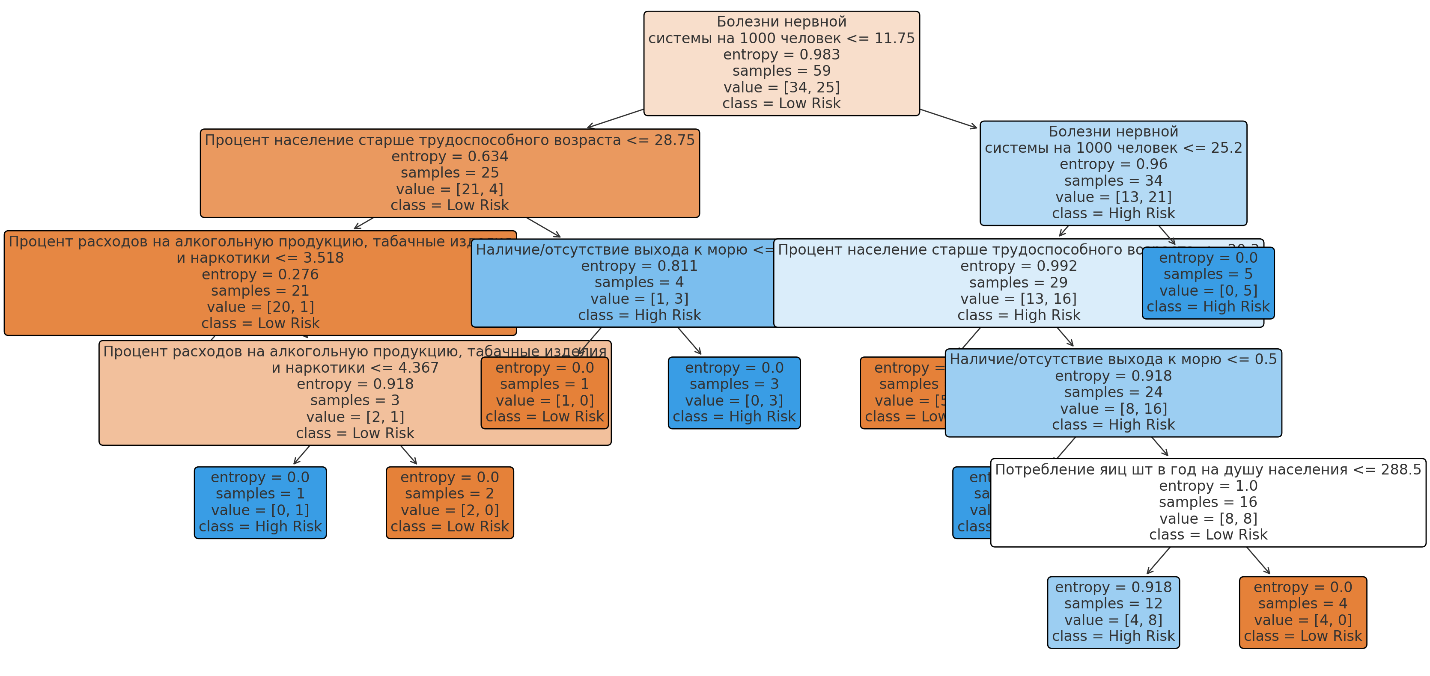


После преобразования переменной мы получили чёткое разделение данных, что позволит точно оценить влияние разных факторов на состояние сердечно-сосудистой системы населения.

**3.2 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.





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.

**3.3 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:

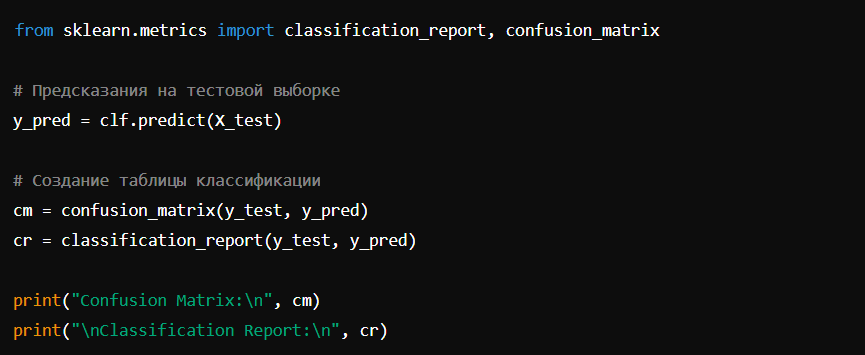
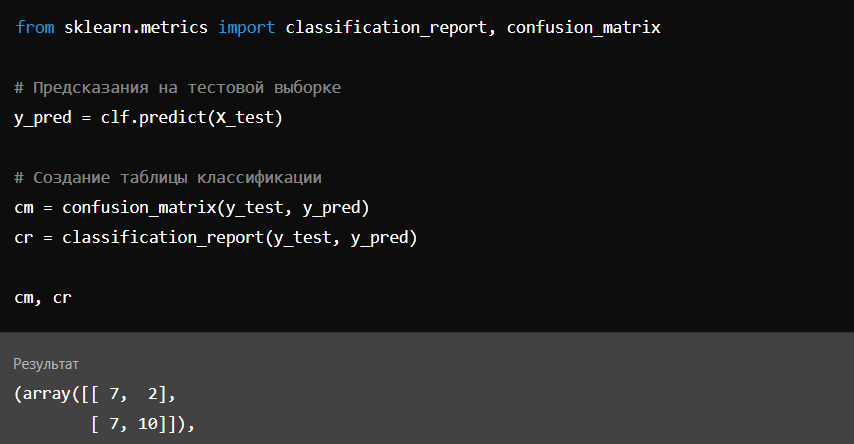


Таблица классификации поможет оценить точность, полноту и F-меру модели, что важно для подтверждения её надёжности и эффективности в прогнозировании риска сердечно-сосудистых заболеваний.



Матрица ошибок:



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.

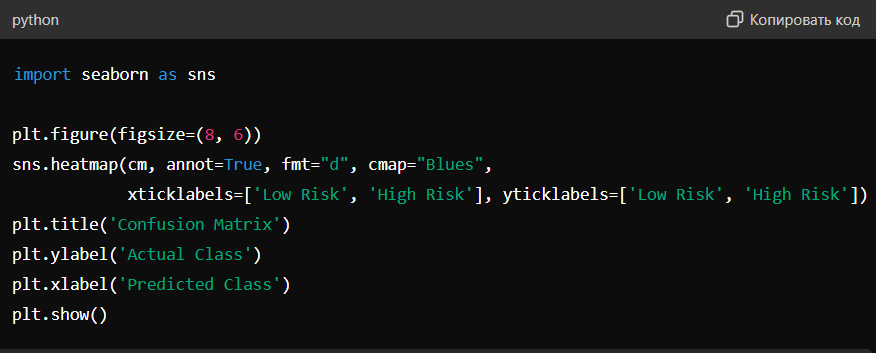
Вывод:

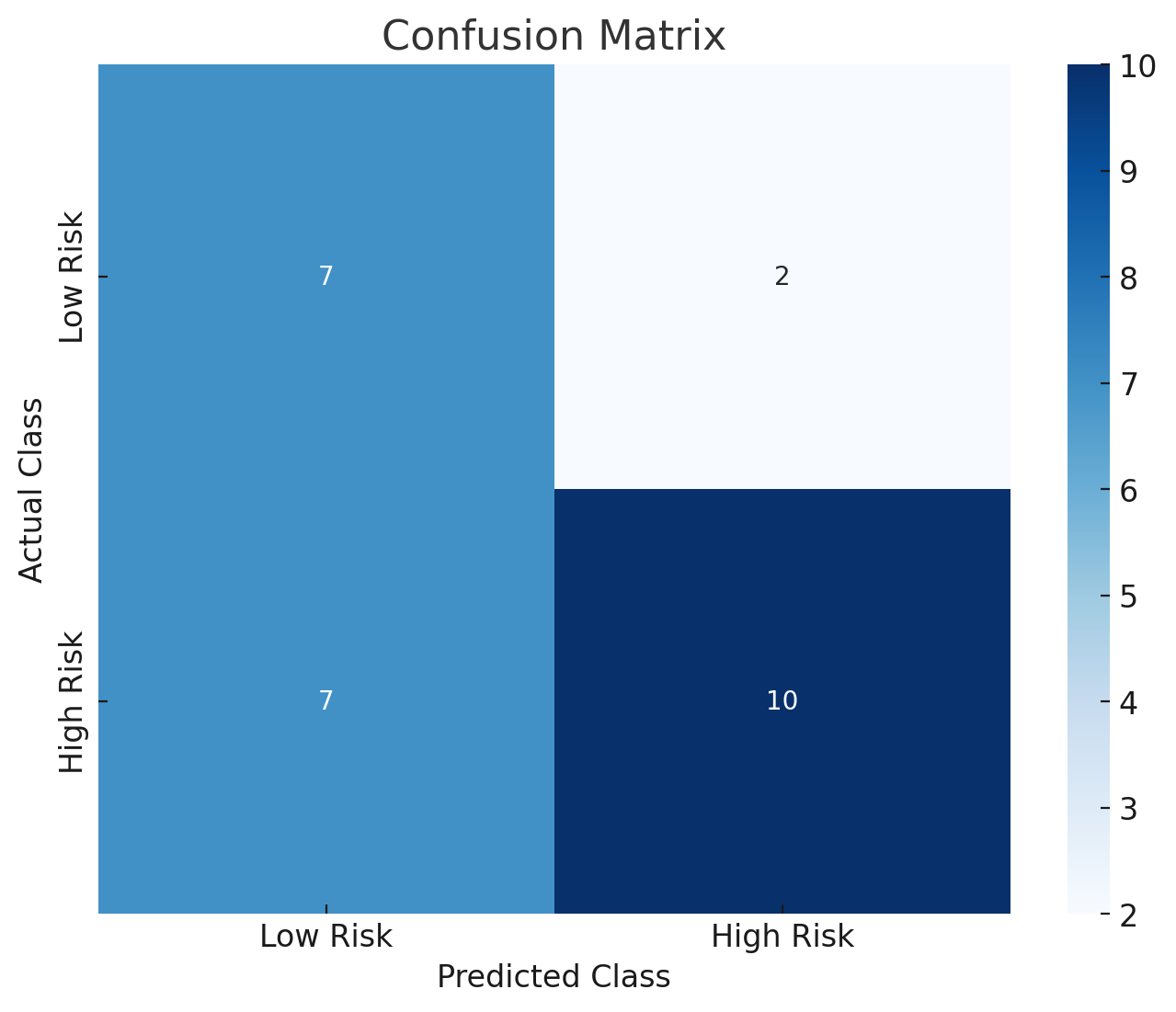
Модель демонстрирует среднюю общую точность (65%). Точность и полнота значительно различаются между двумя классами. В частности, модель лучше предсказывает класс с высоким риском (F-мера 0.69) по сравнению с классом низкого риска (F-мера 0.61). Это указывает на потенциальное переобучение или на недостаточную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

**3.4 Визуализация результатов**

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

Код для визуализации матрицы ошибок:





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

**3.5 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 социоэкономических, экологических и медицинских данных может улучшить предсказательную способность модели.

**4 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.