САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

| Направление | Математика и механика |
| --- | --- |
| Программа | Искусственный интеллект и наука о данных |

Отчёт по практическому заданию

«Реализация байесовской сети доверия»

Выполнил: студент гр. 24.М81-мм, 1 курс

Егоров Ян Игоревич

Преподаватель: Столярова Валерия Фуатовна

Санкт-Петербург, 2024

## 

[**Формулировка задачи 3**](#_4wldv3bqsyjr)

[**Материалы и методы 4**](#_gly5eniu5dl0)

[**Основная часть 5**](#_6cypc2zhm68j)

[**Результаты 9**](#_a18eb8tn71yl)

[**Список использованных источников 10**](#_jky9j7rxnoz5)

## 

## Формулировка задачи

**Основная цель**- предсказывать наличие сердечных заболеваний, по ряду признаков.

**Предметная область**- медицина.

**Описание датасета:**

Сам по себе датасет представляет собой ряд различных признаков, напрямую связанных с рассматриваемым человеком.

Ссылка на датасет- [ml-course/data/heart\_disease/heart.csv at main · rustam-azimov/ml-course](https://github.com/rustam-azimov/ml-course/blob/main/data/heart_disease/heart.csv)

Признаки:

age - возраст в годах

sex - пол (1 = мужчина; 0 = женщина)

cp - тип боли груди

1: типичная стенокардия

2: атипичная стенокардия

3: неангинальная боль

4: бессимптомное течение

trestbps -артериальное давление в состоянии покоя (в мм рт.ст. при поступлении в больницу)

chol -(уровень сахара в крови натощак > 120 мг/дл) (1 = true; 0 = false)

fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg - результаты электрокардиографии в покое

0: нормальное

1: наличие аномалии ST-T

2: показывает гипертрофию левого желудочка

thalach - максимальная частота сердечных сокращений

exang - cтенокардия, вызванная физической нагрузкой (1 = true; 0 = false)

oldpeak - Депрессия ST, вызванная физическими упражнениями, по сравнению с отдыхом

slope - наклон пикового сегмента ST при нагрузке

ca - количество расширенных сосудов (0-3)

thal - тип дефекта

1: нормальный

2: фиксированный дефект

3: обратимый дефект

target - наличие сердечного приступа (1=да, 0=нет).

## Материалы и методы

Непосредственный язык написания программы-Python. Мной были использованы следующие библиотеки: pandas, numpy, pgmpy, matplotlib, seaborn,networkx.

pandas- данная библиотека использовалась для загрузки таблицы и обработки данных внутри нее, работы со столбцами.

numpy- данная библиотека была необходима для использования и применения математических функций.

pgmpy- была использована для построения байесовской сети

pgmpy.models.BayesianModel служит для определения модели (структуры сети).

- pgmpy.models.BayesianNetwork- методы для работы с вероятностными распределениями.

- pgmpy.estimators.HillClimbSearch является алгоритмом для поиска структуры байесовской сети. Он использует жадный подход для оптимизации структуры сети, пробуя различные ребра и выбирая те, которые улучшают оценку по критерию BIC.

- pgmpy.estimators.BicScore — это функция, использующаяся для оценки соответствия между данными и моделью на основе байесовского информационного критерия (BIC).

matplotlib- библиотека использовалась для построения графиков и визуализации данных

seaborn- визуализация данных

networkx- использовалась для построения графов

## 

## 

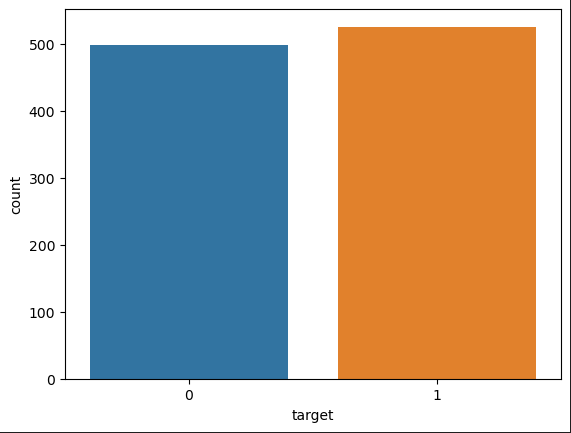
## 

## Основная часть

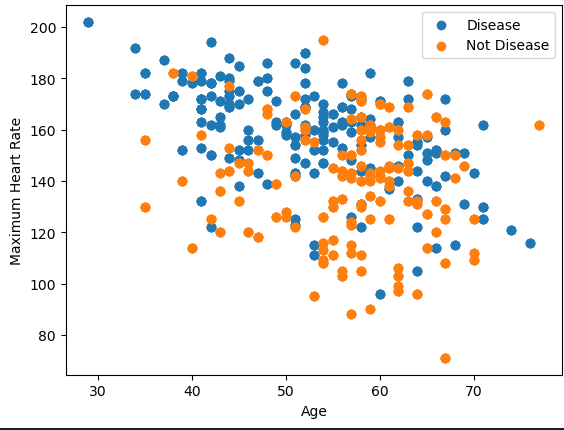
Для начала было необходимо провести разведочный анализ.

Пропущенные значения отсутствовали в выборке.

Для начала была построена гистограмма для целевого признака.

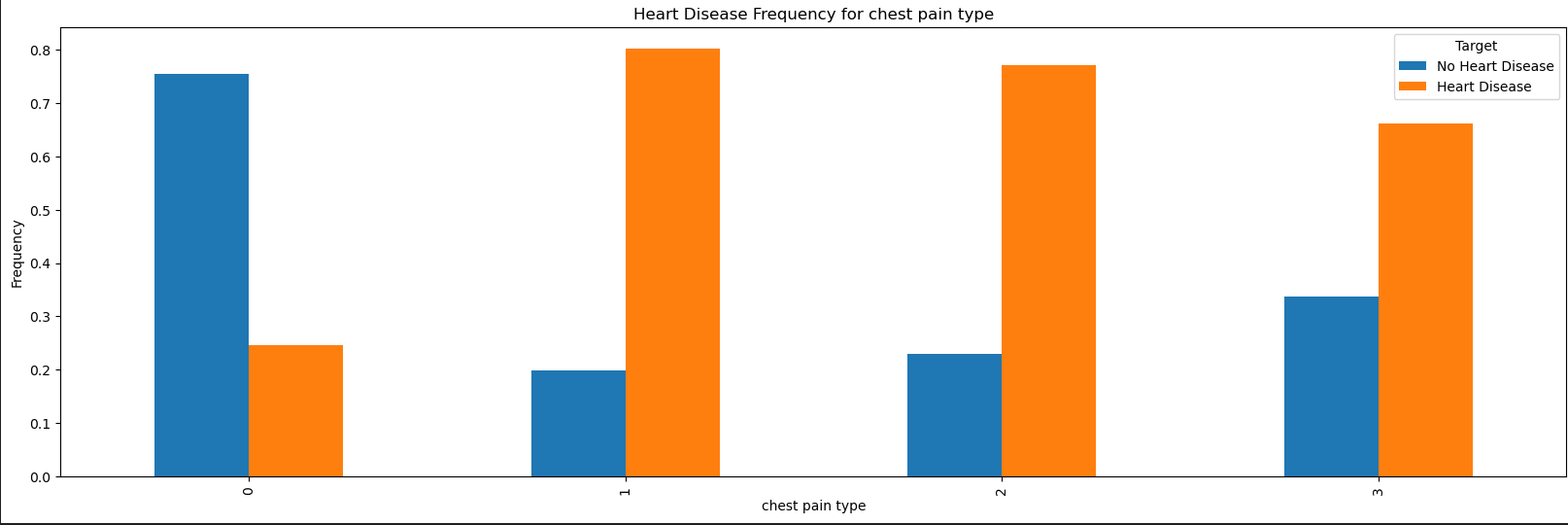


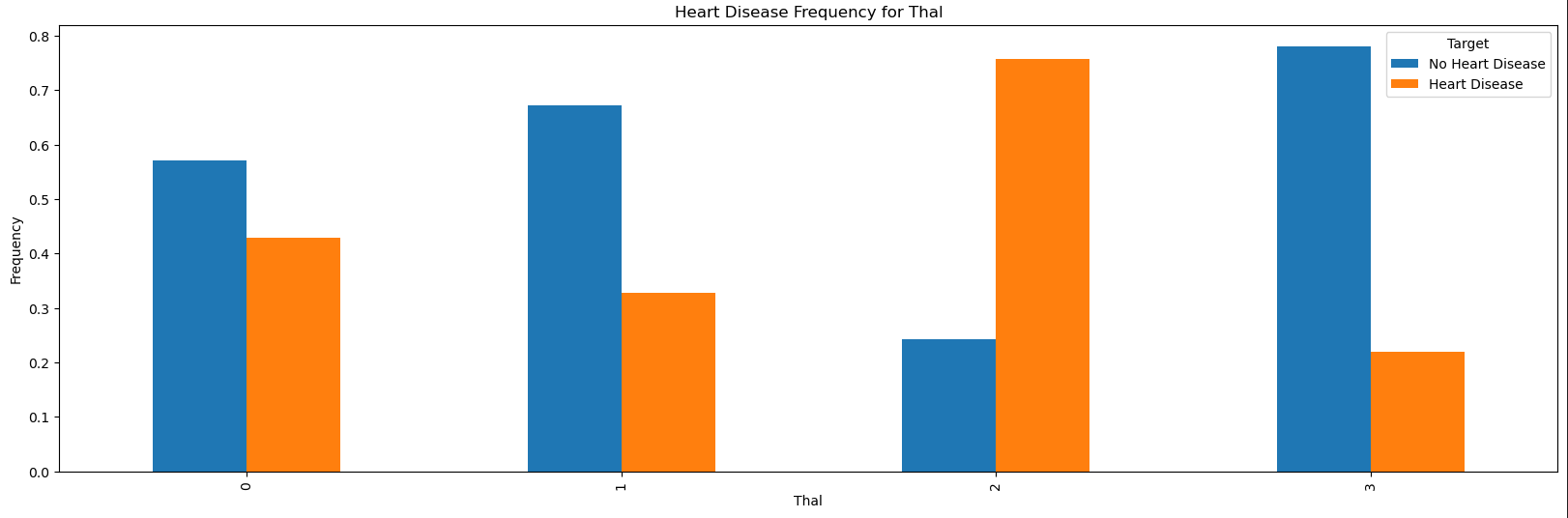
После чего была построена диаграмма для зависимости наличия заболевания от возраста.

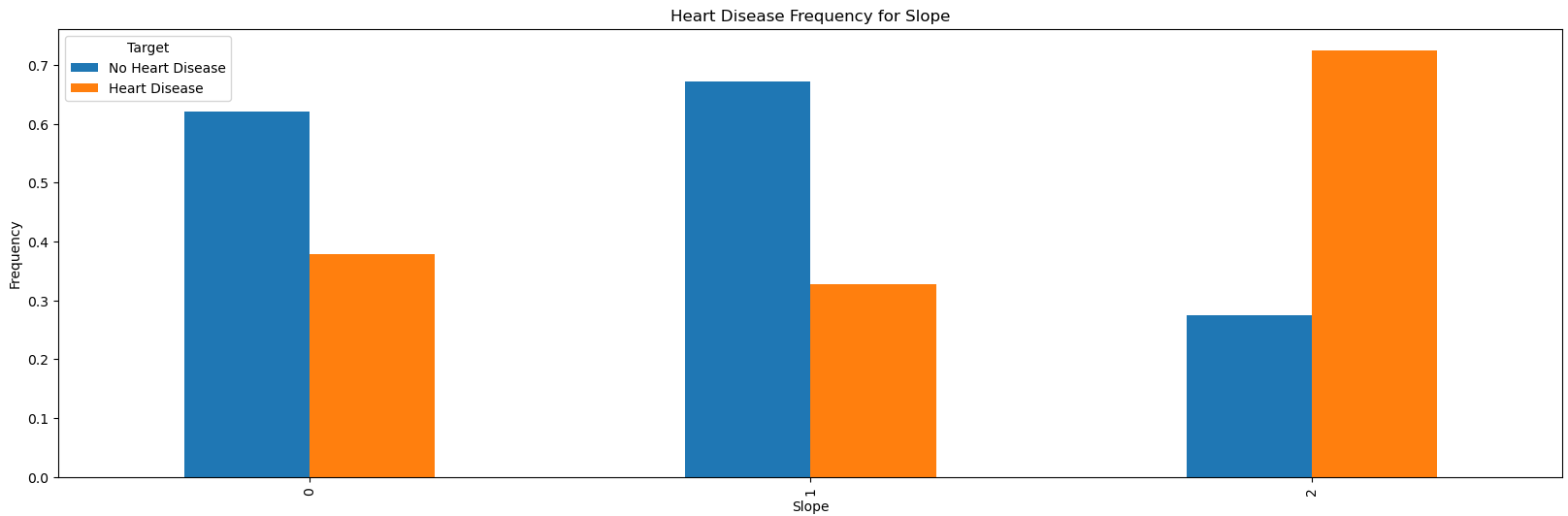


Можно увидеть что с увеличением возраста количество заболеваний увеличивается, в то время как их отсутствие снижается.

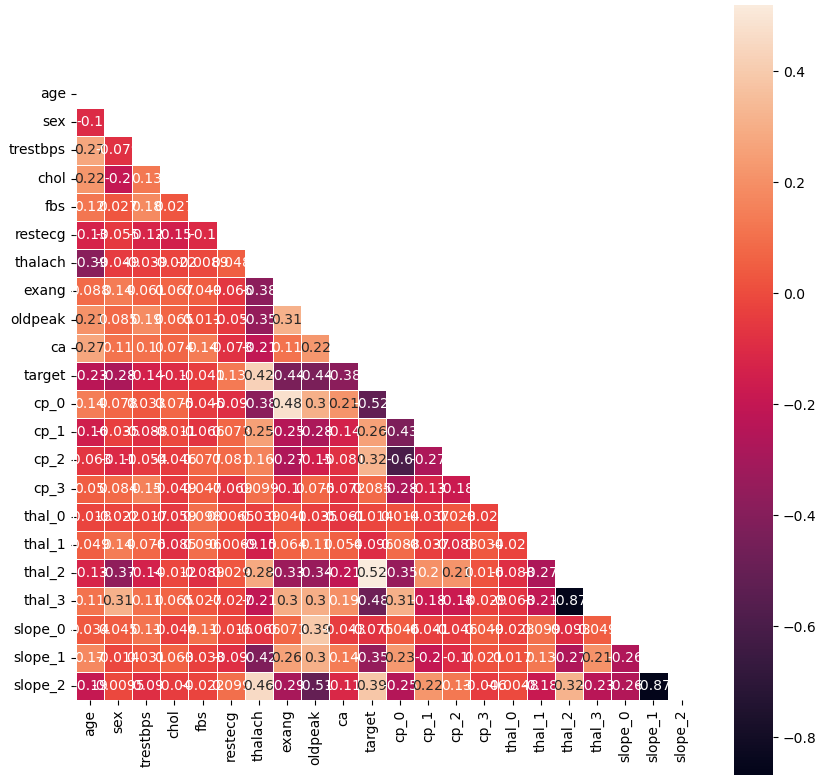
В данной выборке присутствуют категориальные признаки: chest pain type (cp), thal и slope. Для них были построены диаграммы, для того чтобы понять есть ли какая-то однозначная связь с целевым признаком в зависимости от изменения числа внутри этих признаков.





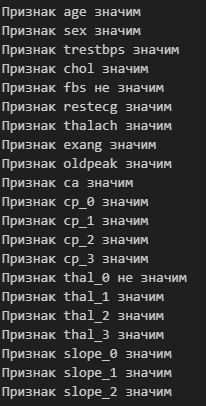
Мы можем увидеть, что однозначная связь отсутствует, поэтому было принято решение перевести данные признаки в дамми.

Также мной была построена корреляционная матрица



Сильная корреляция, >0,7 отсутствует, ее можно увидеть только у бинарных переменных, которые были добавлены для описания одного признака.

После чего также была сделана f статистика для переменных. Были получены следующие результаты:



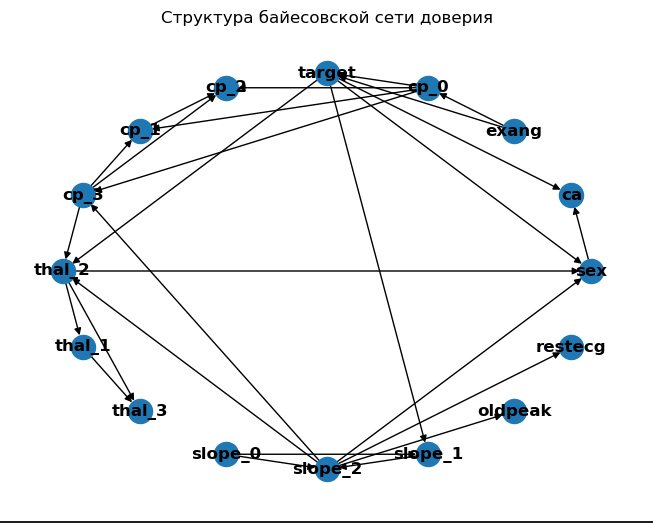
Не значимы оказались признаки thal\_0 и fbs, они были исключены из дальнейшего анализа.

После чего было произведено обучение структуры байесовской сети при помощи метода HillClimbSearch.

Для непосредсвенной оценки использовался критерий BIC (Bayesian Information Criterion, BIC), который рассчитывается по следующей формуле:

, где - максимальное значение функции правдоподобия наблюдаемой выборки с известным числом параметров, - число параметров модели, - объем обучающей выборки.

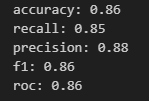
После чего байесовская сеть была построена графически при помощи класса BayesianNetwork. Как итог был получен следующий граф для Байесовской сети.



Как мы можем видеть переменные age, trestbps, chol, thalach были исключены.

Для измерения эффективности модели я буду использовать следующие метрики:

accuracy, recall, precision, f1, roc-auc. Итоговые метрики для всего датасета можно увидеть ниже.



В целом, исходя из полученных метрик, модель обладает довольно неплохим качеством предсказания.

## Результаты

Основной результат данной работы- обучение байесовской сети доверия, которая способна классифицировать наличие сердечных заболеваний.

Для достижения поставленных результатов был проведен разведочный анализ, feature engineering, после чего была создана модель наивной байесовской сети.

Ссылка на GitHub- [BGH612/Bayesian](https://github.com/BGH612/Bayesian)

## 

## 

## 

## 

## 

## 

## 

## Список использованных источников

1. Pandas // pypi.org URL: <https://pypi.org/project/pandas/> (дата обращения: 30.11.2024)
2. Numpy // Numpy.org URL: <https://numpy.org/> (дата обращения: 30.11.2024)
3. Pgmpy// pgmpy.org URL: <https://pgmpy.readthedocs.io/en/latest/> (дата обращения: 30.11.2024)
4. Matplotlib// matplotlib.org URL: <https://matplotlib.org/> (дата обращения: 30.11.2024)
5. Seaborn// seaborn.pydata.org URL: <https://seaborn.pydata.org/> (дата обращения: 30.11.2024)
6. Networkx// networkx.org URL: <https://networkx.org/> (дата обращения: 30.11.2024)