****Московский государственный технический университет им. Н.Э. Баумана  
Кафедра «Системы обработки информации и управления»

Домашнее задание №1  
по дисциплине  
«Методы машинного обучения»

Выполнил:  
студент группы ИУ5-21М  
Ян Тяньци

Москва — 2022 г.

**1. Задание**

Домашнее задание по дисциплине направлено на решение комплексной задачи машинного обучения с учителем. Домашнее задание включает выполнение следующих шагов:

1.Поиск и выбор набора данных для построения модели машинного обучения. На основе выбранного набора данных строится модель машинного обучения для решения или задачи классификации, или задачи регрессии.

2.Для выбранного датасета (датасетов) на основе материалов лекций, в целях улучшения выборки, решить следующие задачи (если это необходимо в данном датасете):

устранение пропусков в данных;

кодирование категориальных признаков;

нормализацию числовых признаков;

масштабирование признаков;

обработку выбросов для числовых признаков;

обработку нестандартных признаков (которые не является числовым или категориальным);

отбор признаков, наиболее подходящих для построения модели;

устранение дисбаланса классов в случае решения задачи классификации на дисбалансированной выборке.

3.Обучить модель и оценить метрики качества для двух выборок :

исходная выборка, которая содержит только минимальную предобработку данных, необходимую для построения модели (например, кодирование категориальных признаков).

улучшенная выборка, полученная в результате полной предобработки данных в пункте 2.

4.Построить модель с использованием произвольной библиотеки AutoML.

5.Сравнить метрики для трех полученных моделей.

**2. Ход выполнения работы**

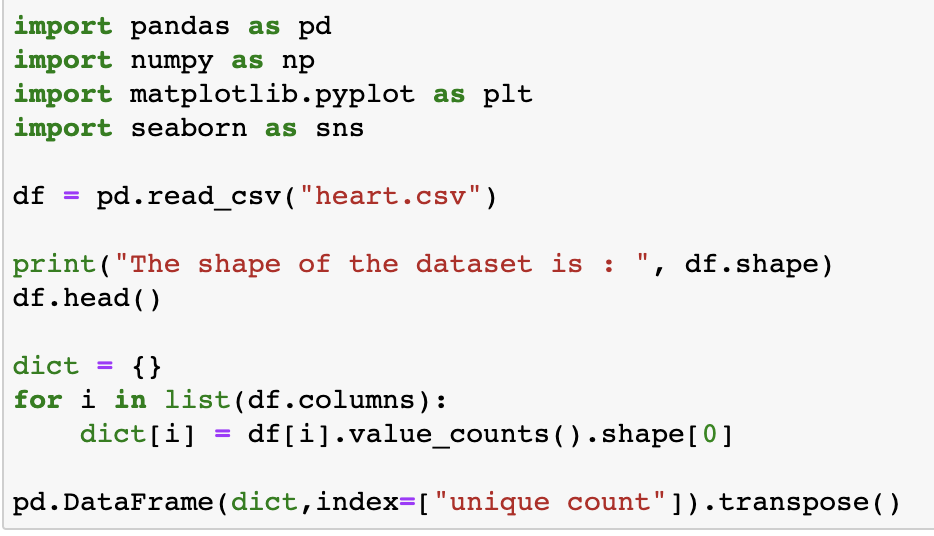
**2.1. Выбор набора данных**

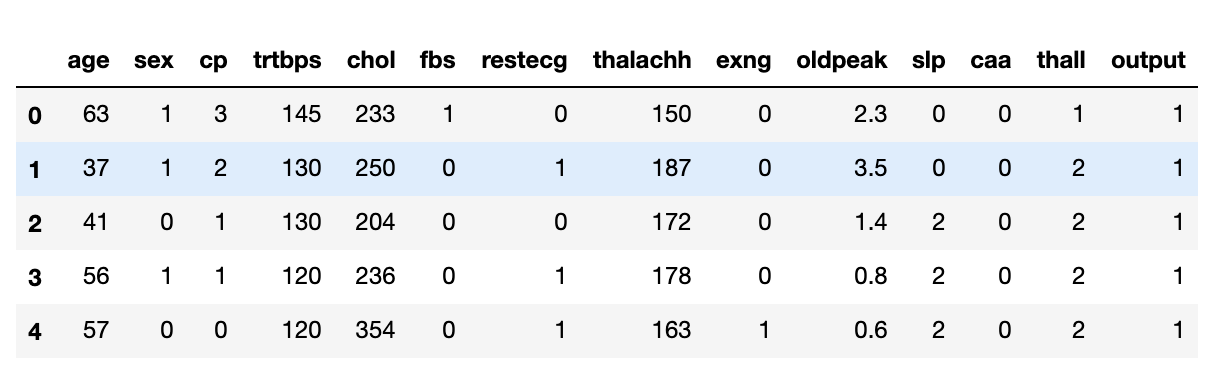
Heart Attack Analysis & Prediction Dataset

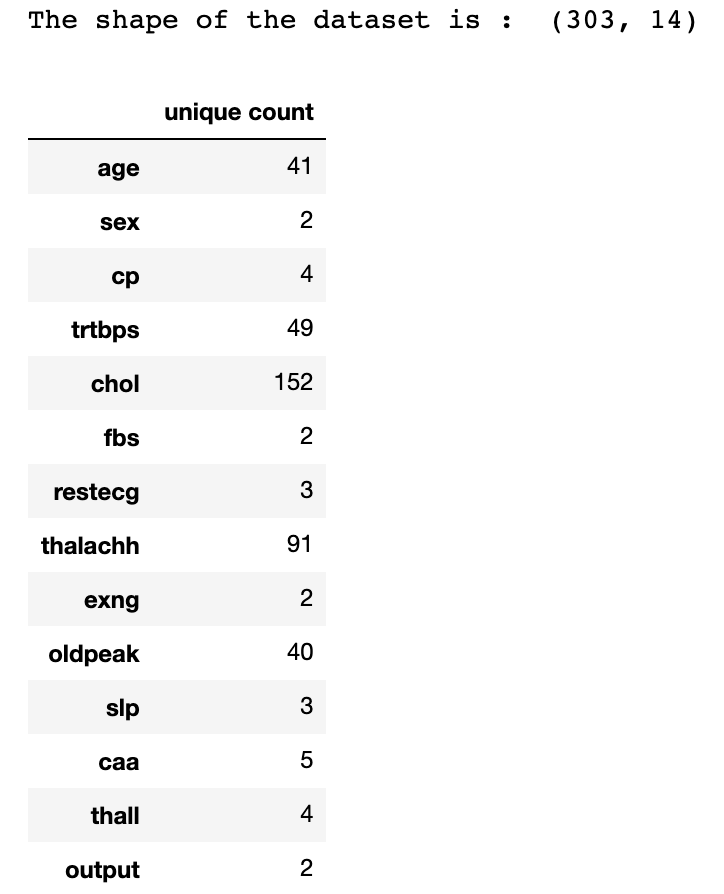
A dataset for heart attack classification

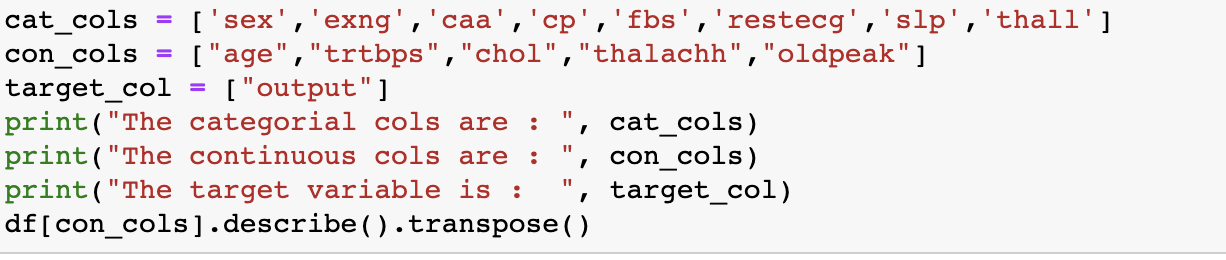
Набор данных для классификации сердечных приступов

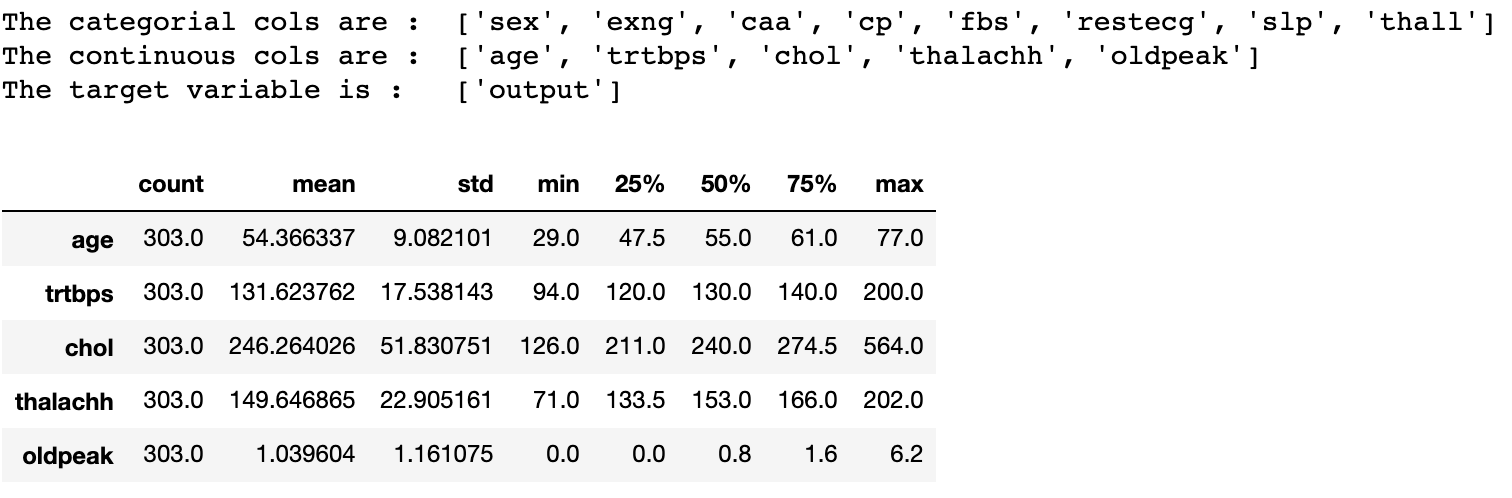
**2.2. Проведение разведочного анализа данных**

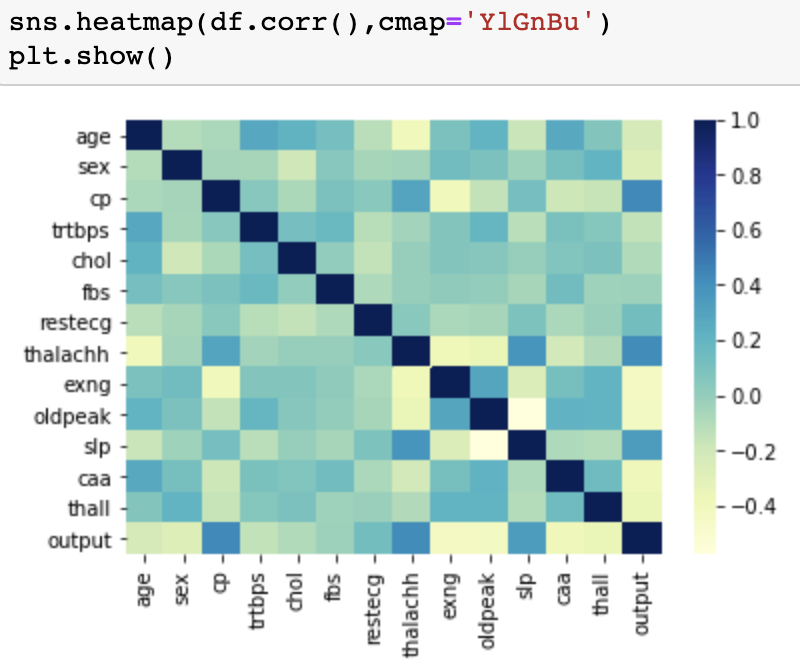
****

****

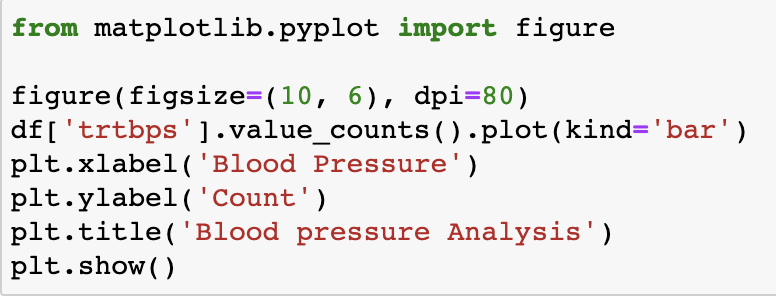
****

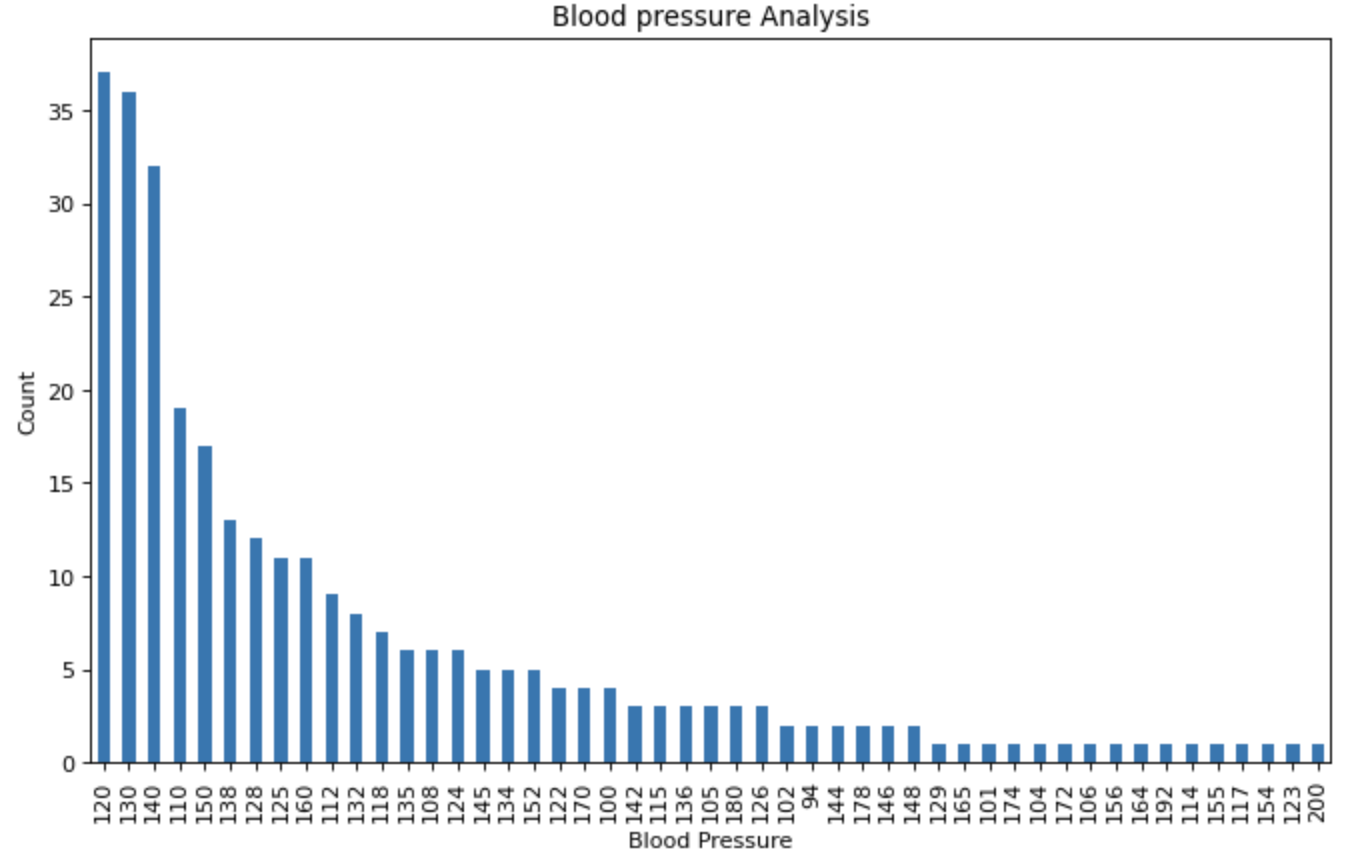
****

****

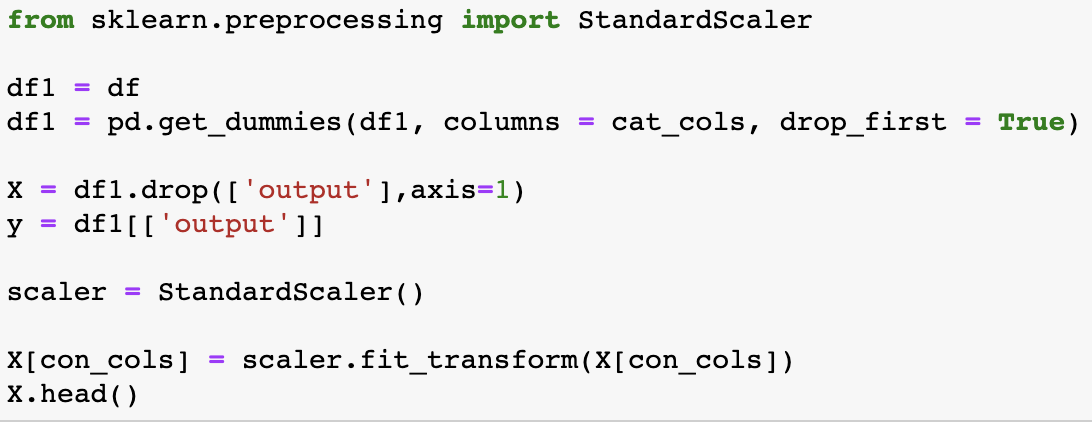
****

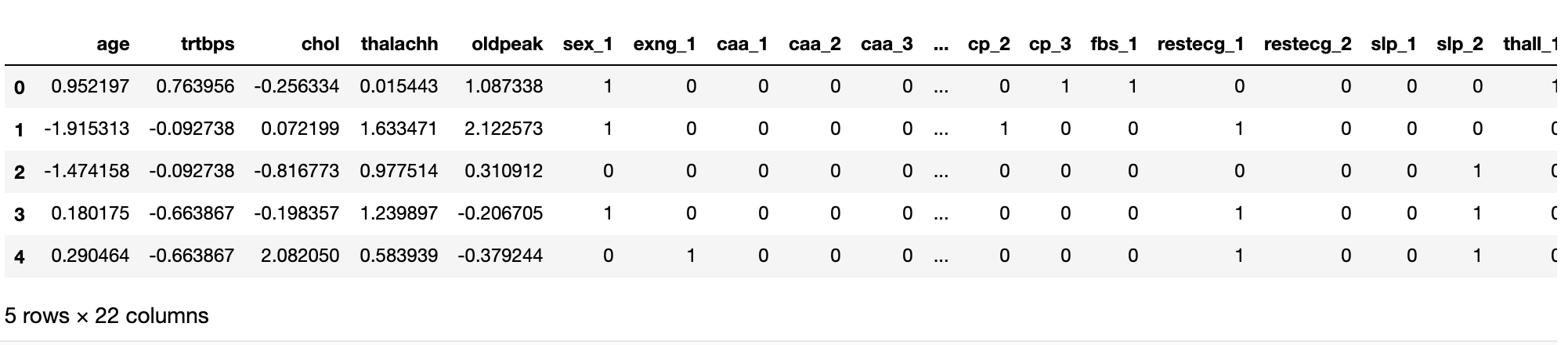
У кого бы ни была боль в груди 2-го типа, возможна высокая вероятность сердечного приступа.



****

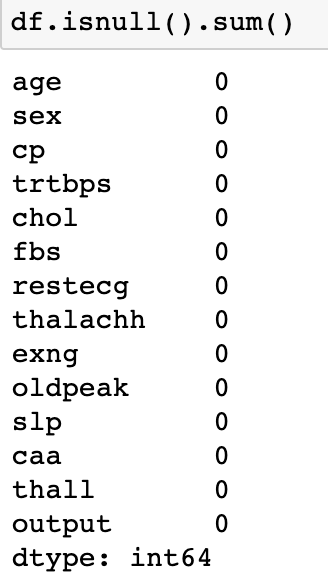
**2.3** **исходная выборка, которая содержит только минимальную предобработку данных, необходимую для построения модели**

****

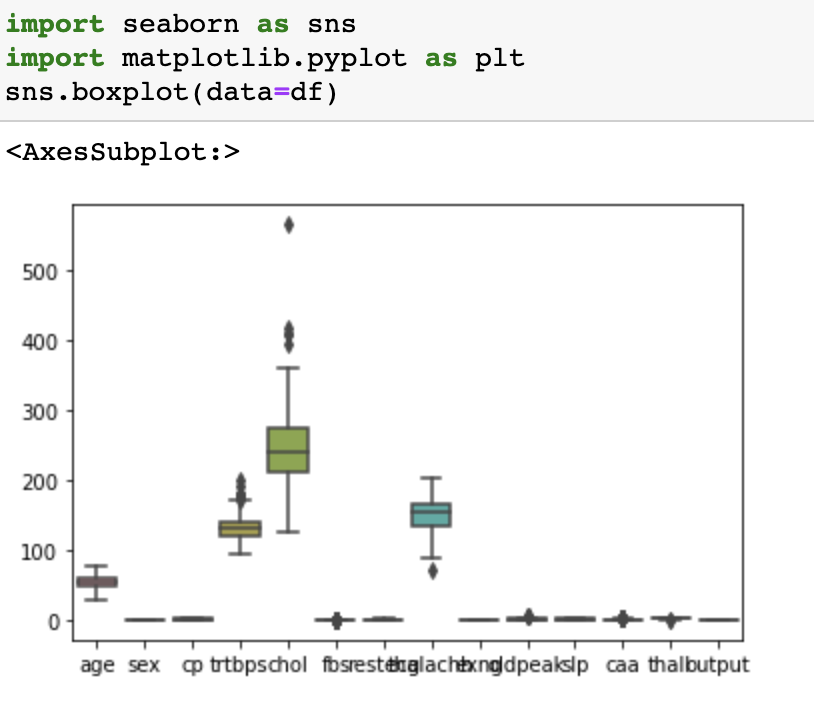
****

**2.3 у****лучшенная выборка, полученная в результате полной предобработки данных в пункте 2.**

2.3.1 Проверьте, является ли это нулевым значением

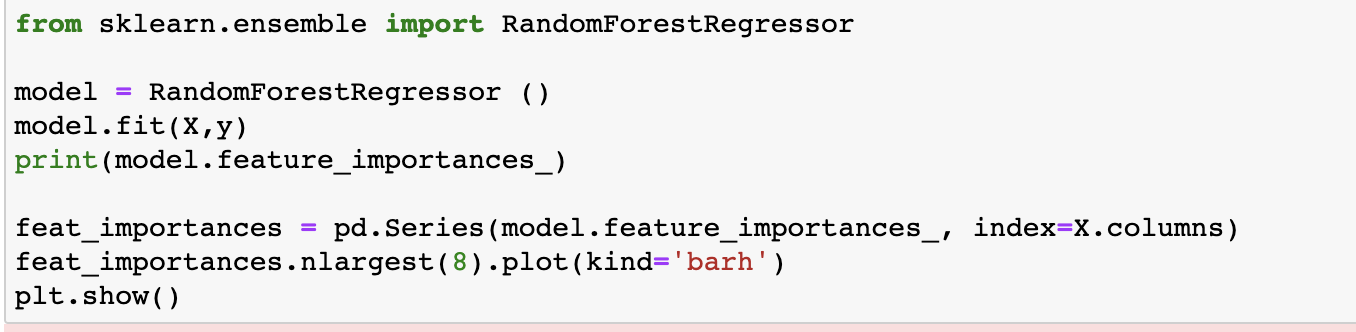
****

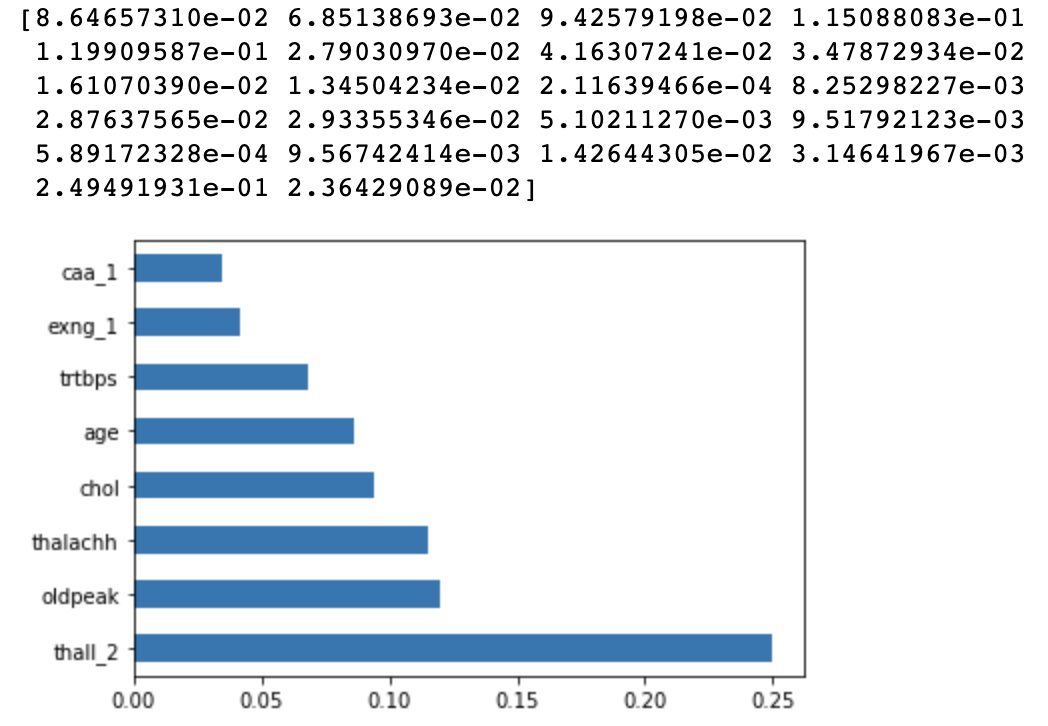
2.3.2 Нарисуйте прямоугольную диаграмму, чтобы проверить, есть ли выбросы.

****

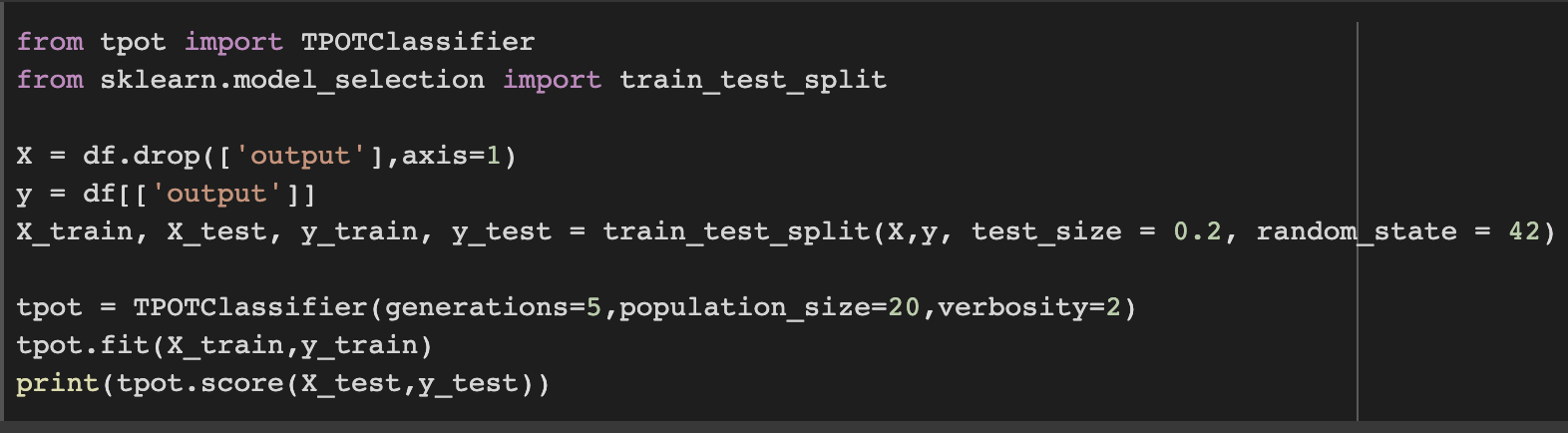
В этом наборе данных собраны все данные конкретных значений.Выбросы для наборов данных, скорее всего, существуют в реальности.Так что выбросы не обрабатываются.

2.3.3 С помощью метода случайного леса выберите 8 индикаторов с наибольшим весом..

****

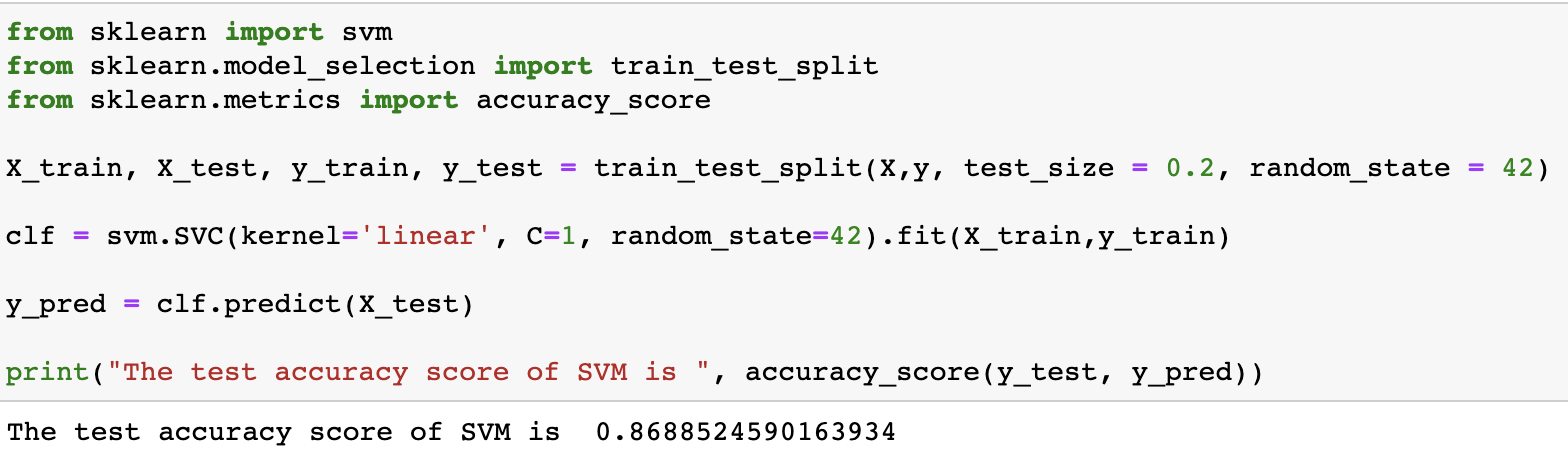
****

**2.4 Построить модель с использованием произвольной библиотеки AutoML.(TPOT)**

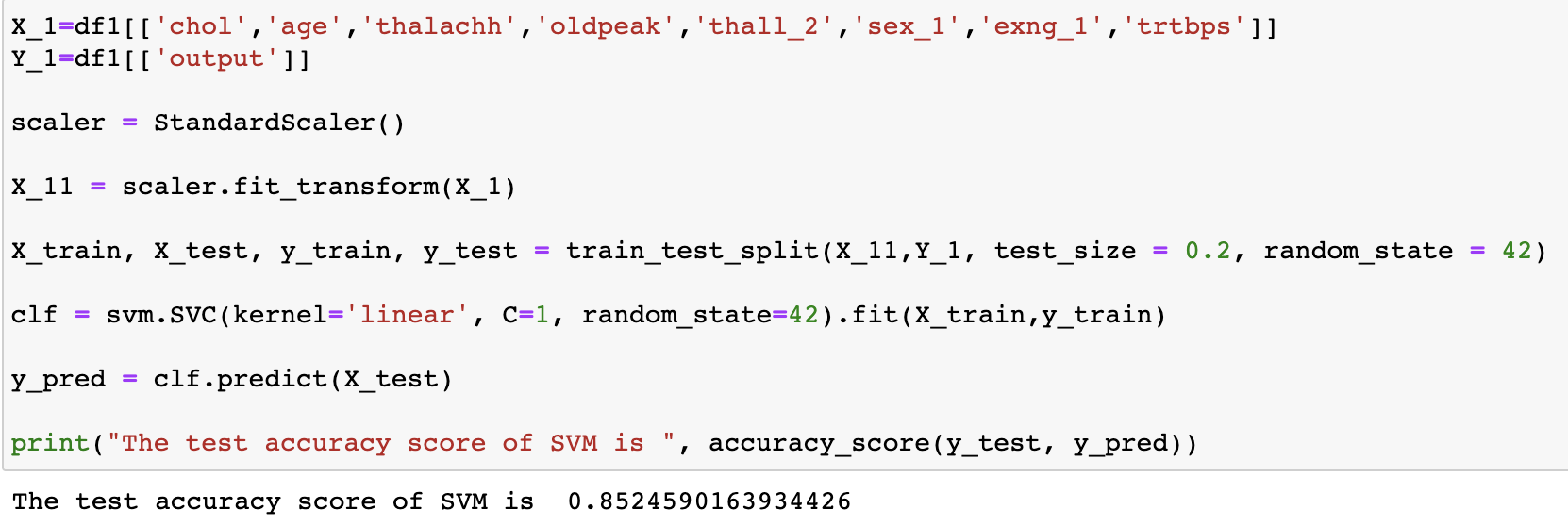
****

**2.4 Сравнить метрики для трех полученных моделей.**

**2.4.1**исходная выборка, которая содержит только минимальную предобработку данных, необходимую для построения модели (результат)

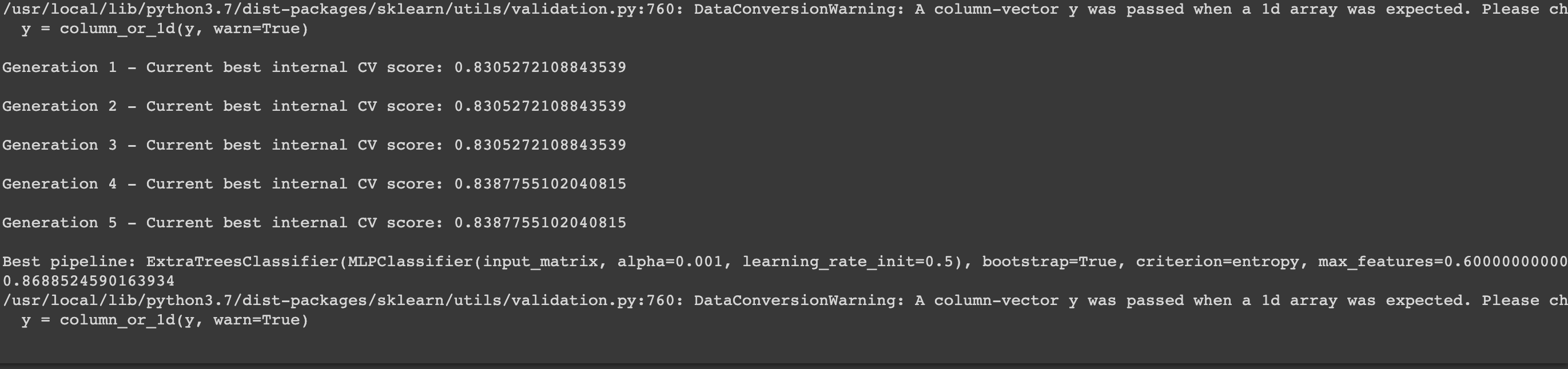


**2.4.2** лучшенная выборка, полученная в результате полной предобработки данных в пункте 2 (результат)



Почему точность падает после обработки данных? После метода AUTOML я обнаружил, что самая высокая точность составляет 86,88%, такая же точность, как и при отсутствии обработки данных. Когда это возможно, поскольку количество данных и индикаторов в этом наборе данных невелико, все данные более важны.

**2.4.3** AutoML(результат)



**Список литературы**

[1] Гапанюк Ю. Е. Домашнее задание по дисциплине «Методы машинного обучения»[Электронный ресурс] // GitHub. –– 2019. –– Режим доступа: https://github.com/ugapanyuk/ml\_course/wiki/MMO\_DZ (дата обращения: 06.05.2019).

[2] You are my Sunshine [Electronic resource] // Space Apps Challenge. — 2017. Access mode: https://2017.spaceappschallenge.org/challenges/earth-and-us/

you-are-my-sunshine/details (online; accessed: 22.02.2019).

[3] dronio. Solar Radiation Prediction [Electronic resource] // Kaggle. — 2017. — Access mode: https://www.kaggle.com/dronio/SolarEnergy (online; accessed: 18.02.2019).

[4] Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource] //Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/stable/ (online; accessed: 20.02.2019).

[5] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).

[6] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode:http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).

[7] Chrétien M. Convert datetime.time to seconds [Electronic resource] // Stack Overflow. 2017. — Access mode: https://stackoverflow.com/a/44823381 (online; accessed:20.02.2019).