Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»

*Факультет Программной инженерии и компьютерной техники*

# Лабораторная работа №4-6

Модуль 2. Методы МО

Группа: P33312

Выполнил:
Скориков Родион Викторович

Проверил:
Кугаевских Александр Владимирович

Санкт-Петербург
2023г

# Содержание

Введение ..... 3
Линейная регрессия ..... 4
Используемый датасет ..... 4
Исследование датасета ..... 4
Тренировка и оценка моделей ..... 7
Вывод ..... 7
Метод KNN ..... 9
Используемый датасет ..... 9
Нормализация ..... 9
Тренировка и оценка моделей ..... **Ошибка! Закладка не определена.**
Вывод ..... 10
Деревья решений ..... 12
Используемый датасет ..... 12
Очистка датасета ..... 12
Визуализация данных ..... 12
Тренировка и оценка модели ..... 13
Вывод ..... 15
Вывод ..... 16

# Введение

Целью данного модуля является изучение простейших методов машинного обучения на основе наборов датасетов. Для каждого датасета используется разный метод, который подходит под набор данных

# Линейная регрессия

## Используемый датасет

В данной работе был использован датасет про жилье калифорнии.

**Параметры:**

Longitude - Мера того, как далеко на запад находится дом; более отрицательное значение дальше на запад

Latitude - Мера того, как далеко на севере находится дом; более высокое значение находится дальше на север

Housing Median Age - Средний возраст дома в квартале; меньшее число - более новое здание

Total Rooms - Общее количество комнат в блоке

Total Bedrooms - Общее количество спален в блоке

Population - Общее количество людей, проживающих в блоке

Households - Общее количество домохозяйств, группа людей, проживающих в жилой единице, для квартала

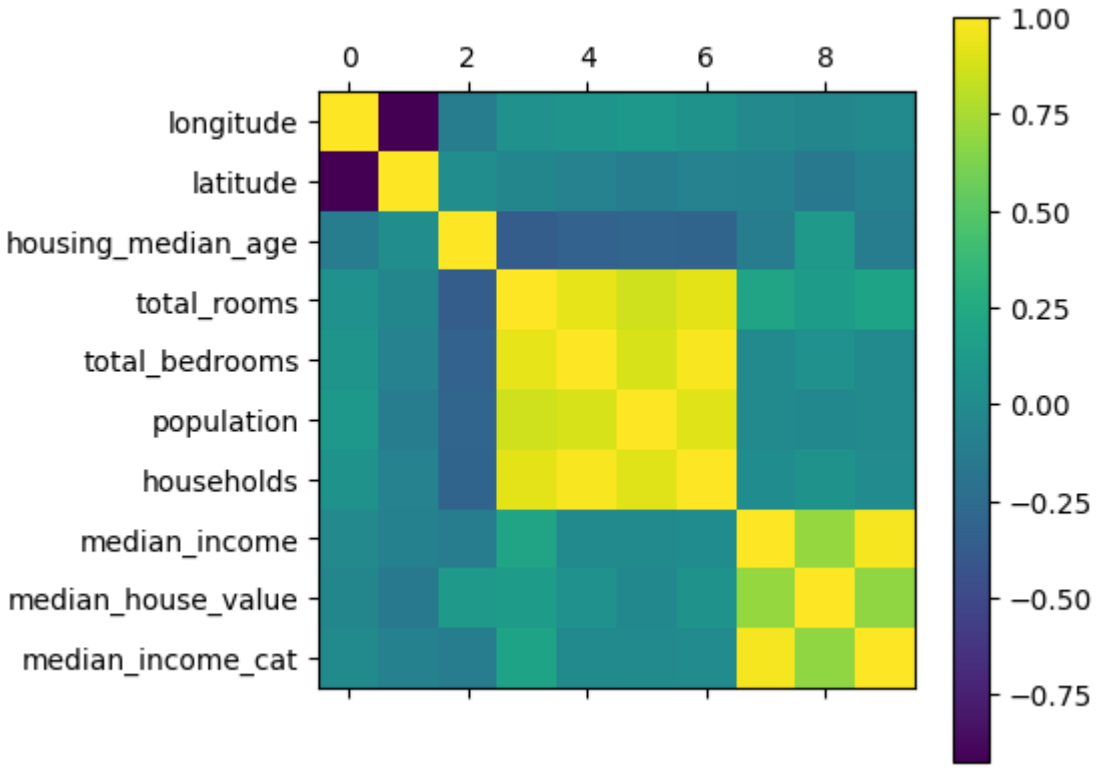
Median Income - Средний доход домохозяйств в многоквартирном доме (измеряется в десятках тысяч долларов США)

**Целевая переменная:**

Median House Value - Средняя стоимость дома для домохозяйств в квартале (измеряется в долларах США)

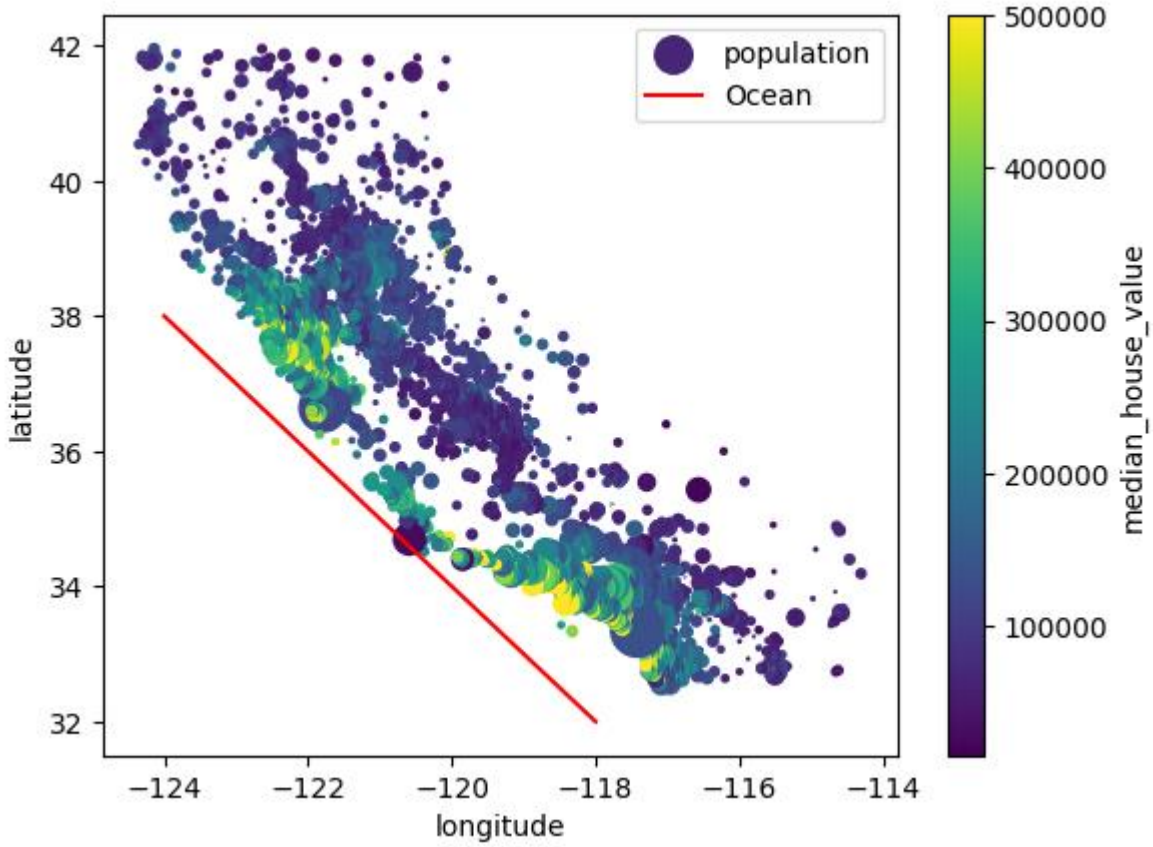
## Исследование датасета

Во первых, проверили корреляцию между признаками:

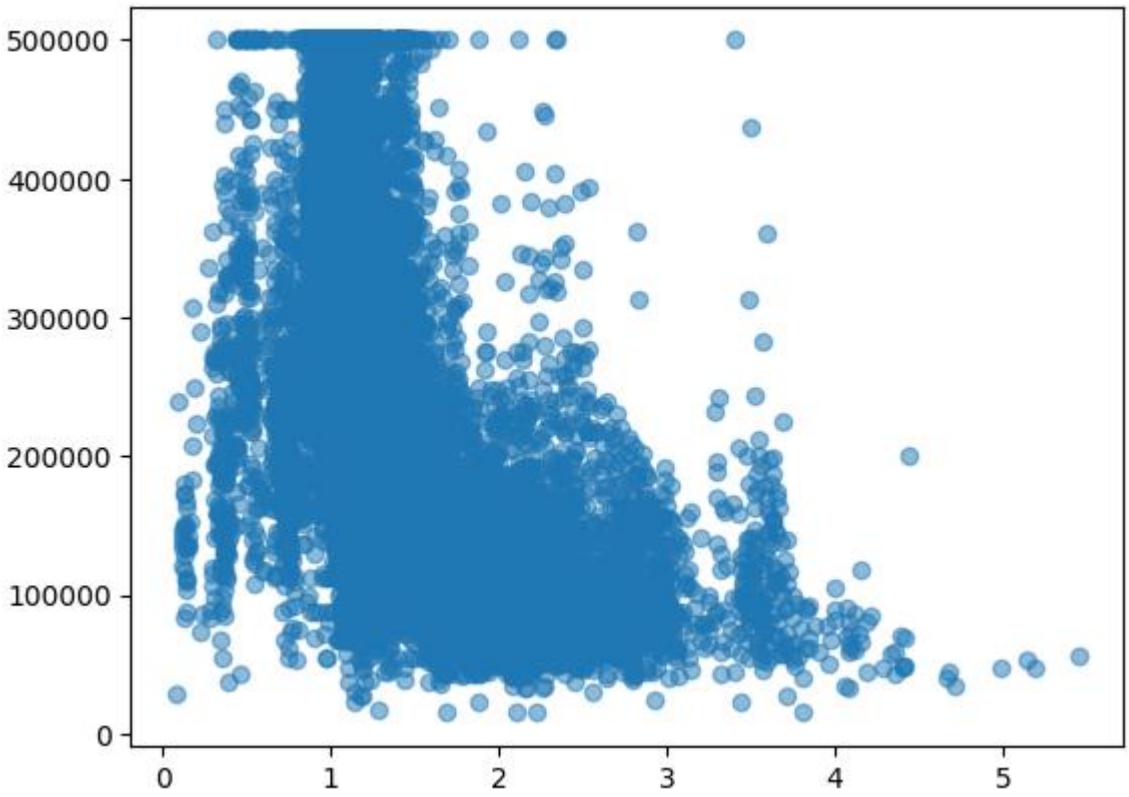



Были обнаружены коллинеарные признаки Households и Total Bedrooms, было принято решение удалить признак Total Bedrooms, поскольку для линейной регрессии будет выгоднее избавиться от коллинеарных признаков.

Далее, чтобы сгенерировать более полезных признак, чем Longitude и Latitude, был введен признак «расстояние до океана».




И был построен график зависимости Median House Value от этого нового признака, видно, что действительно есть зависимость.




## Тренировка и оценка моделей

Для тренировки и тестирования модели, исходный датасет был разделён на две партии: 60% для тренировки, 40% для тестирования.

Для модели был выбран метод наименьших квадратов, который по формуле вычисляет необходимые коэффициенты.

Затем, было проведены тренировки трех моделей с использованием разных наборов признаков, и для оценки этих моделей была использована метрика «Коэффициент детерминации».

Первая модель по всем признакам дала очень плохую метрику, поскольку для линейной регрессии следует избежать коллинеарные признаки.

Вторая модель по признаку Median Income, дала метрику 0,48.

И третья модель по Median Income, Total Rooms и Distance to Ocean дала метрику 0,59.

## Вывод

Поскольку была произведена не очень качественная очистка, а нормализация данных отсутствовала, метрики были не очень хорошие, но по результату получившихся моделей, можно сделать вывод, что в Калифорнии цена дома

наиболее сильно зависит от дохода и близости к океану. Если хотите дешево жить, то нужно переезжать в гетто и подальше от океана.

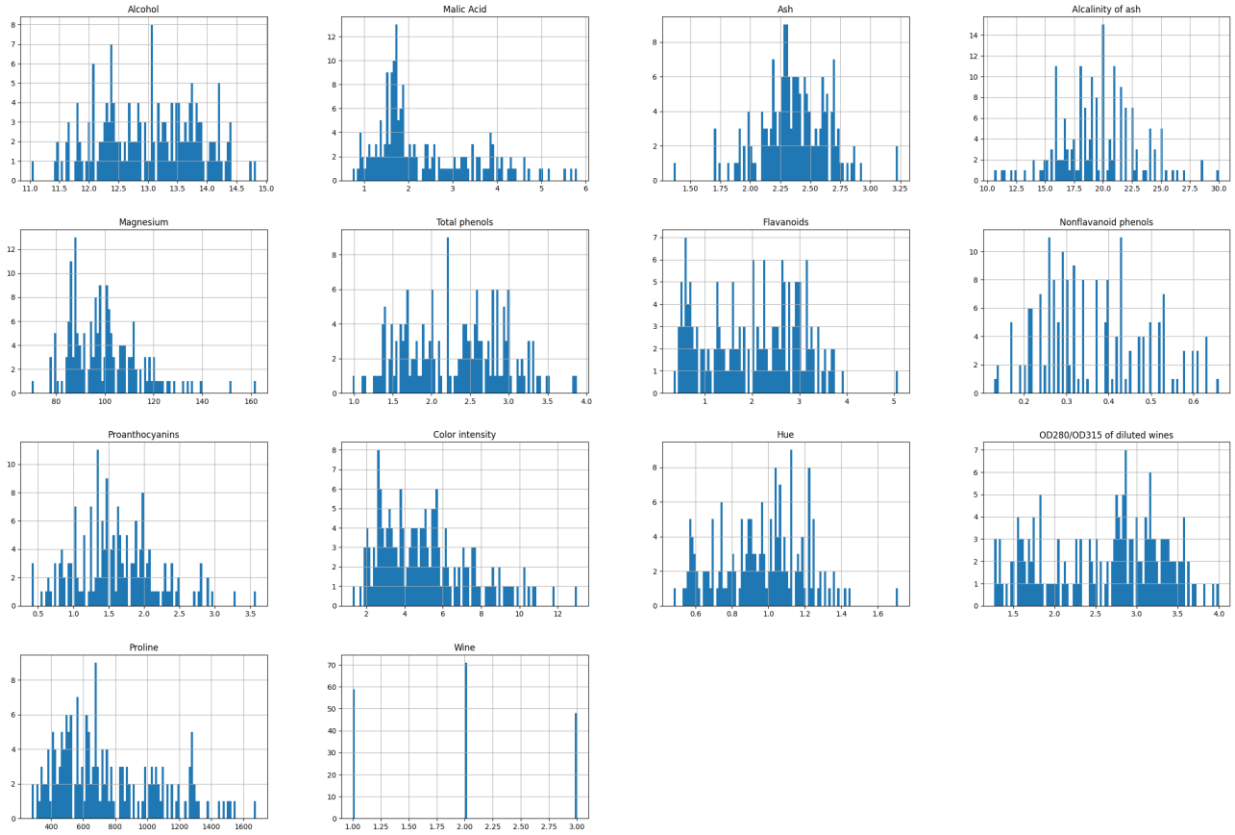
# Метод KNN

## Используемый датасет

Для данной работы был использован датасет по параметрам вина, из которых нужно вывести категорию вина. Были даны 13 численных признаков, и 1 категориальный целевой признак.

## Нормализация

Изначально было рассмотрено распределение всех признаков




Поскольку большинство признаков более менее равномерно распределены, было принято решение использовать MinMax нормализацию.

## Тренировка и оценка моделей

Для тренировки модели было использовано 60% исходного датасета.

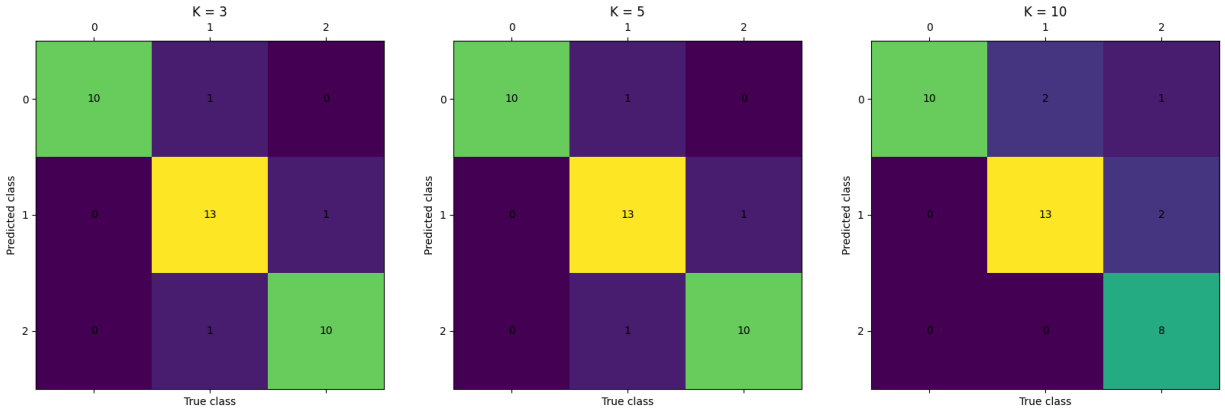
Поскольку метод KNN не требует никакой тренировки, данные просто сохраняются в модели для дальнейшего использования при предсказывании категории вина.

Для оценки качества моделей была выбрана метрика матрицы ошибок, поскольку предсказуемый признак является категориальным.

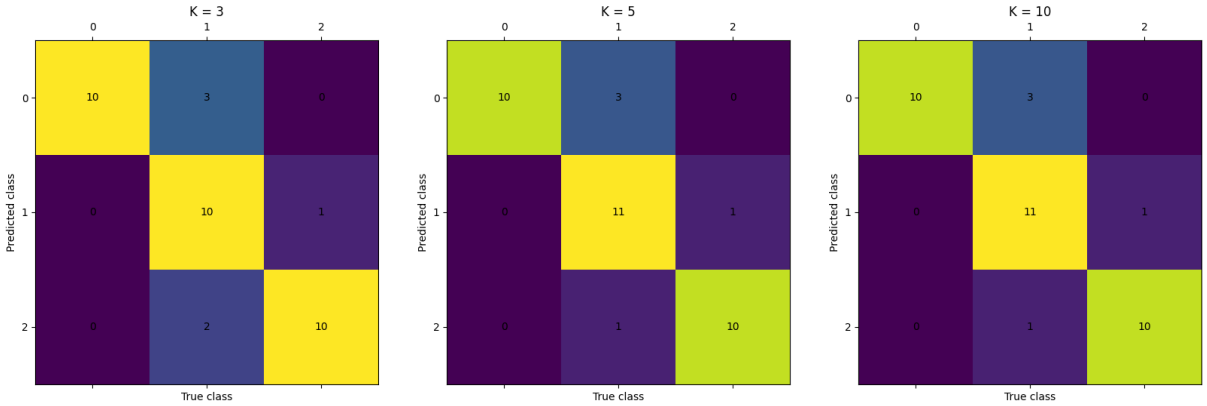
Для первой модели, в котором случайным образом были выбраны признаки

- `Total phenols`
- `Magnesium`
- `Flavanoids`
- `Hue`
- `Proline`
- `Alcohol`
- `Proanthocyanins`
- `Nonflavanoid phenols`

Были получены такие метрики при разных значениях K




Для второй модели, в которой были использованы все признаки, были получены следующие матрицы:




## Вывод

Метод KNN следует использовать, когда недостаточно ресурсов для тренировки модели, и модели должна выдавать категориальные признаки. Но один недостаток — это потенциально долгое предсказывание, поскольку необходимо совершить сортировку, для получения ближайших соседей.

В данной работе можно было ещё провести эксперименты с использованием разных функций расстояний.

# Деревья решений

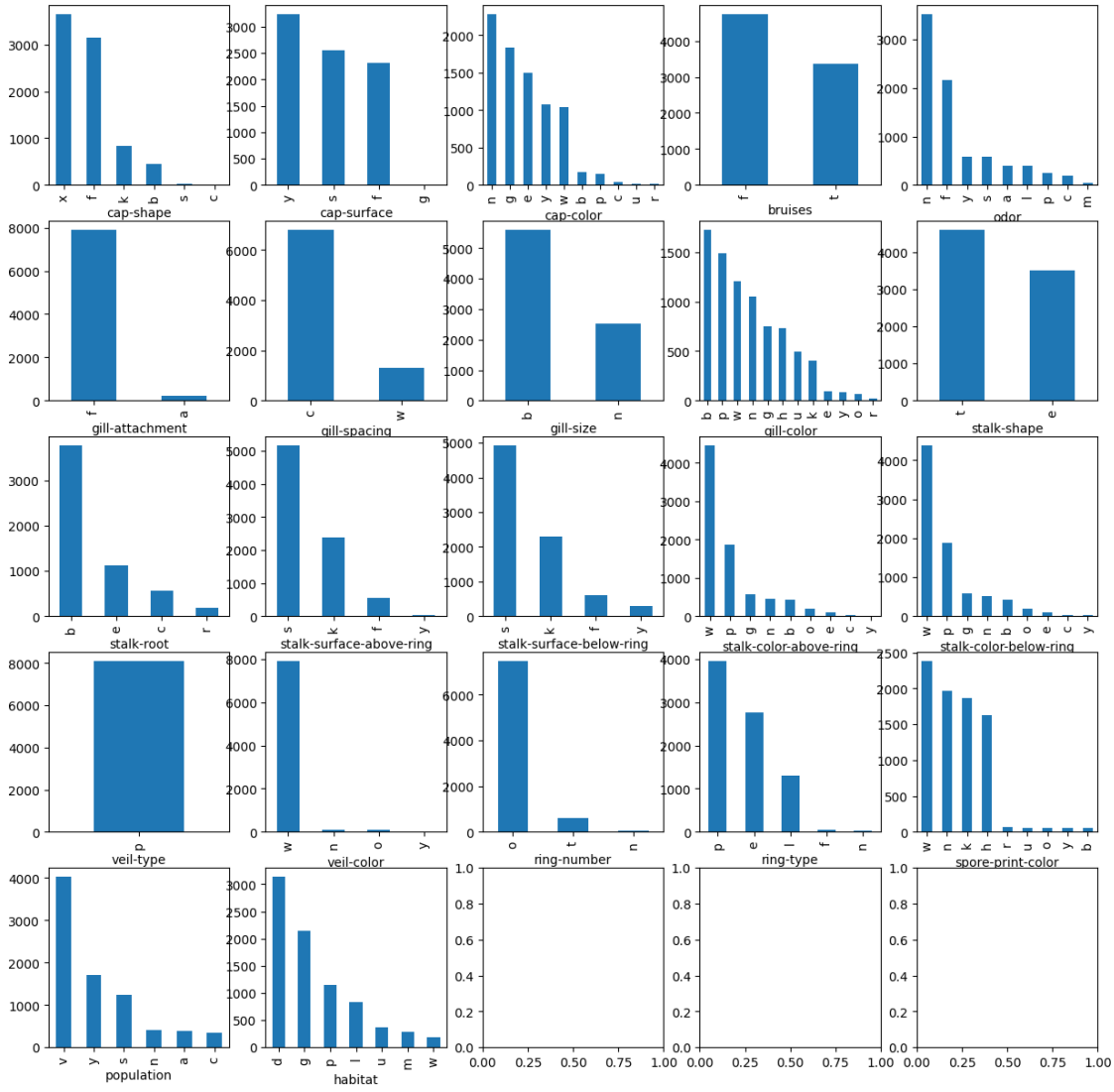
## Используемый датасет

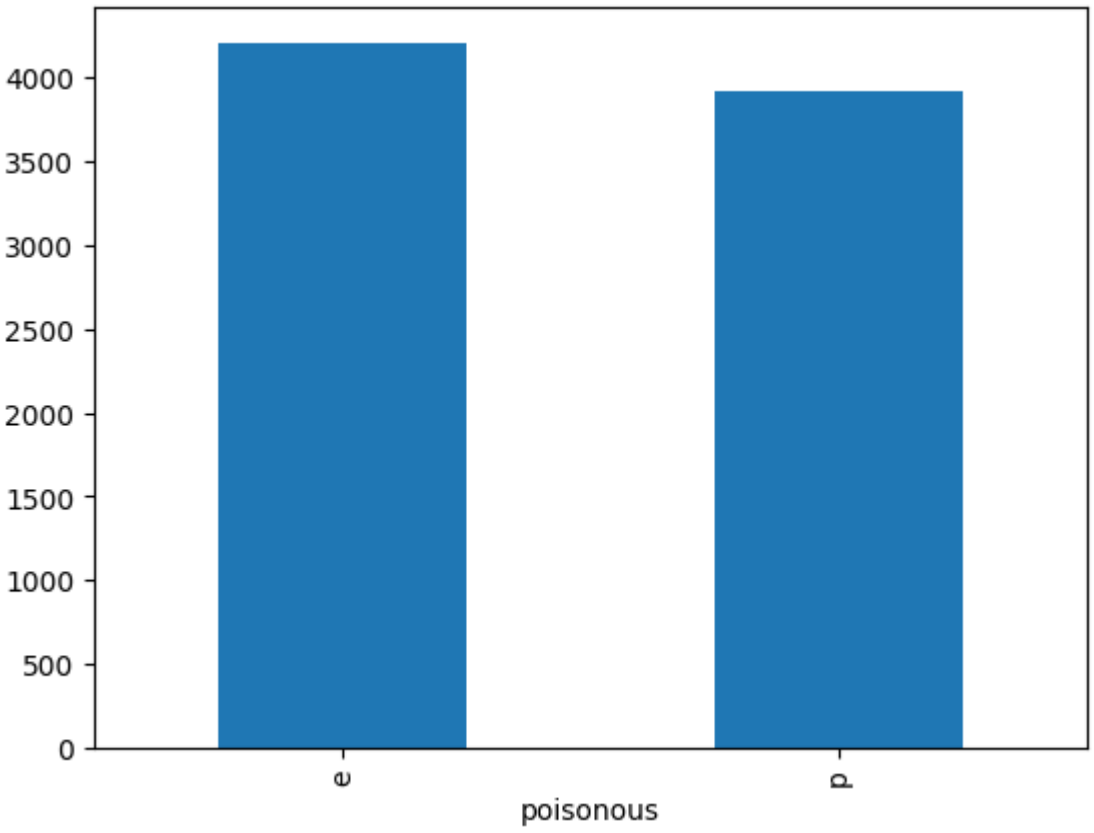
Для данной работы был использован датасеты о съедобности грибов. В датасете все признаки являются категориальными. Тем самым нет необходимости совершать нормализацию данных.

## Очистка датасета

В признаке Stalk Root было обнаружено значительное количество отсутствующих данных, которые были заполнены модой этого признака.

## Визуализация данных






## Тренировка и оценка модели

Для тренировки было выделено 70% исходного датасета.

Также отобраны sqrt(n) признаков в датасете.

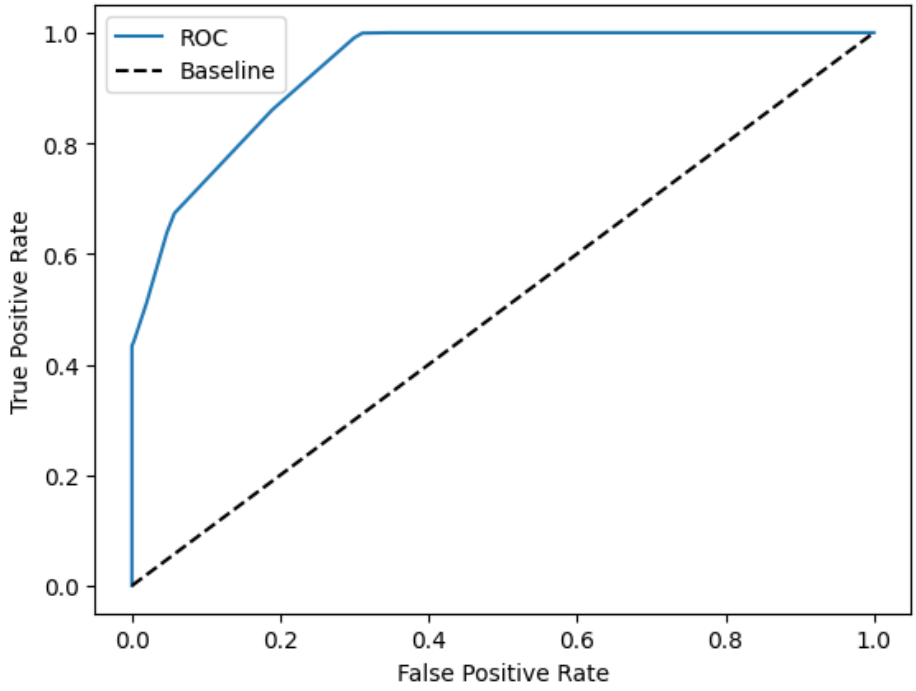
Использовались признаки: cap-color stalk-color-below-ring population ring-type stalk-shape.

Для оценки модели, мы использовали E (edible) как положительный признак, а P (poisonous) как отрицательный, и были использованы такие метрики как матрица запутанности, accuracy, precision, recall и fallout.

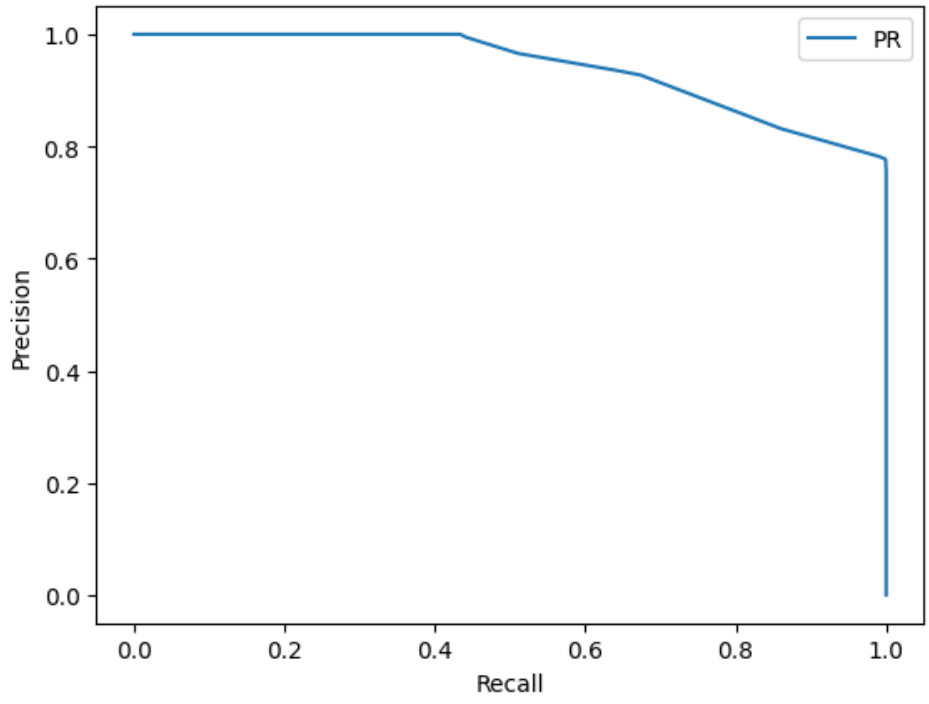
| | Is P | Is E |
|---|---|---|
| Predicted P | 1019 | 33 |
| Predicted E | 150 | 1235 |

Accuracy: 0.92 Precision: 0.89 Recall: 0.97

AUC-ROC:




AUC-PR



## Вывод

Метод деревьев решений подходит для данных, в котором все признаки являются категориальными. При этом нужно аккуратно подходить к тренировке моделей, поскольку она может при тренировки вырастить очень длинные ветки.

## Вывод

В данном модуле были рассмотрены различные методы машинного обучения, каждый из которых имеет свою область применения, поскольку методы имеют разную производительность, результаты и требования к данным. Поэтому следует тщательно изучить методы, чтобы знать, какой метод необходимо применять для разных случаев жизни.