Московский авиационный институт
(Национальный исследовательский университет)
Институт информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Лабораторная работа № 1

по курсу «Искусственный интеллект»

Студент: Шухова А.В.
Группа: М8О-307Б-17
Преподаватель: Самир Халид
Дата:
Оценка:
Подпись:

Москва, 2020

Постановка задачи

Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять собой табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять собой набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

Выбранные датасеты:

Japan Hostel Dataset (https://www.kaggle.com/koki25ando/hostel-world-dataset)
Mushroom Classification (https://www.kaggle.com/uciml/mushroom-classification)

Japan Hostel Dataset

Описание входных данных

- hostel.name – название отеля
- City – название города, в котором находится отель
- price.from – минимальная цена за одну ночь проживания
- Distance – расстояние от центра города (км)
- summary.score – суммарный балл оценок
- rating.band – рейтинговая группа
- atmosphere – рейтинговая оценка атмосферы
- cleanliness – рейтинговая оценка чистоты
- facilities – рейтинговая оценка объектов
- location.y – рейтинговая оценка местоположения
- security – рейтинговая оценка безопасности
- staff – рейтинговая оценка персонала
- valueformoney – рейтинговая оценка стоимости
- lon – долгота
- lat – широта

Анализ данных

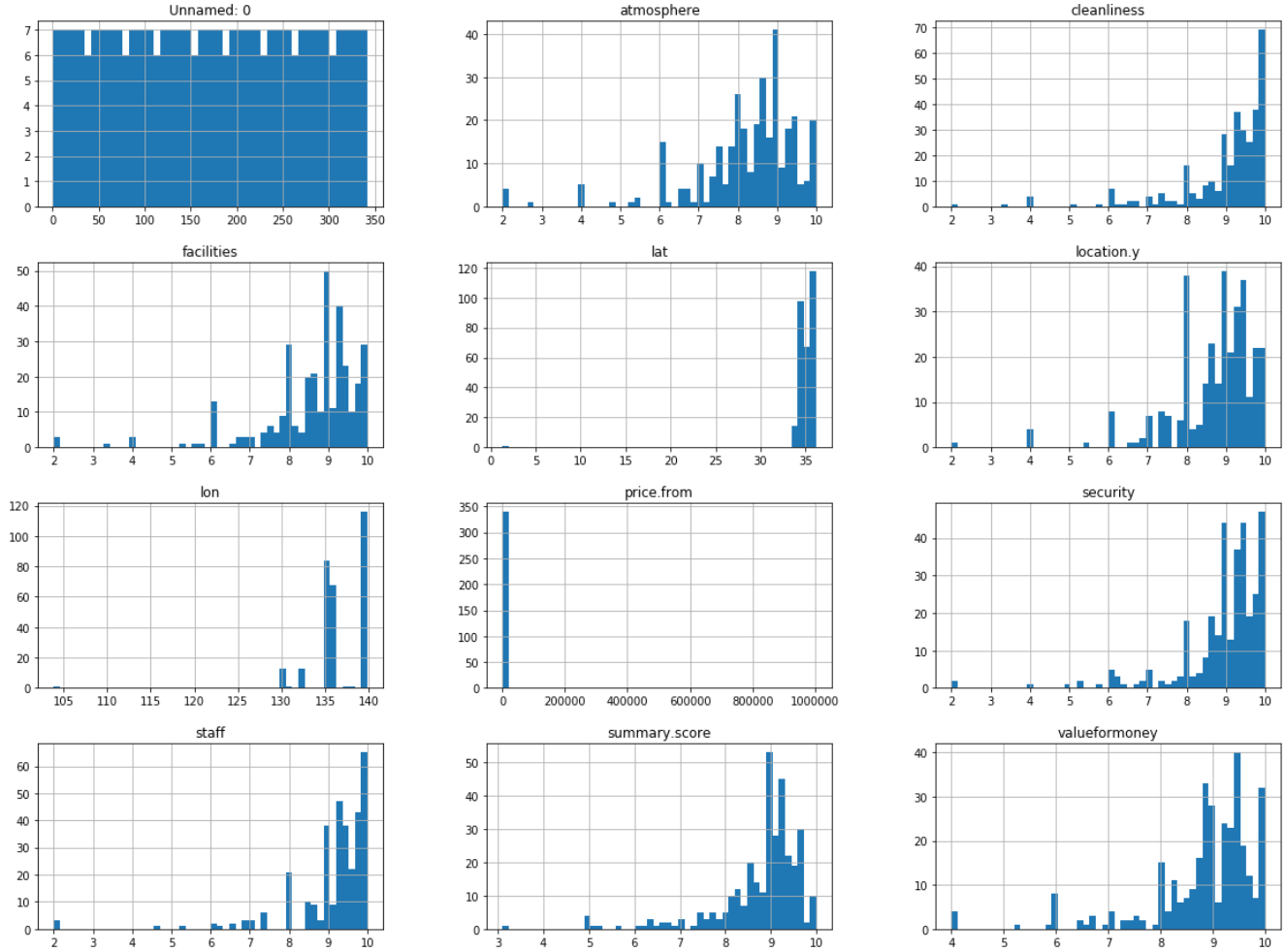
Типы признаков

- Категориальные признаки: hostel.name, City, rating.band
- Количественные признаки: price.from, summary.score, atmosphere, cleanliness, facilities, location.y, security, staff, valueformoney, lon, lat

Размер

- Строк: 342
- столбцов: 16

Распределение признаков с числовыми полями



Решаемая задача

Предсказание признака summary.score.

Признаки, выбранные для решения задачи

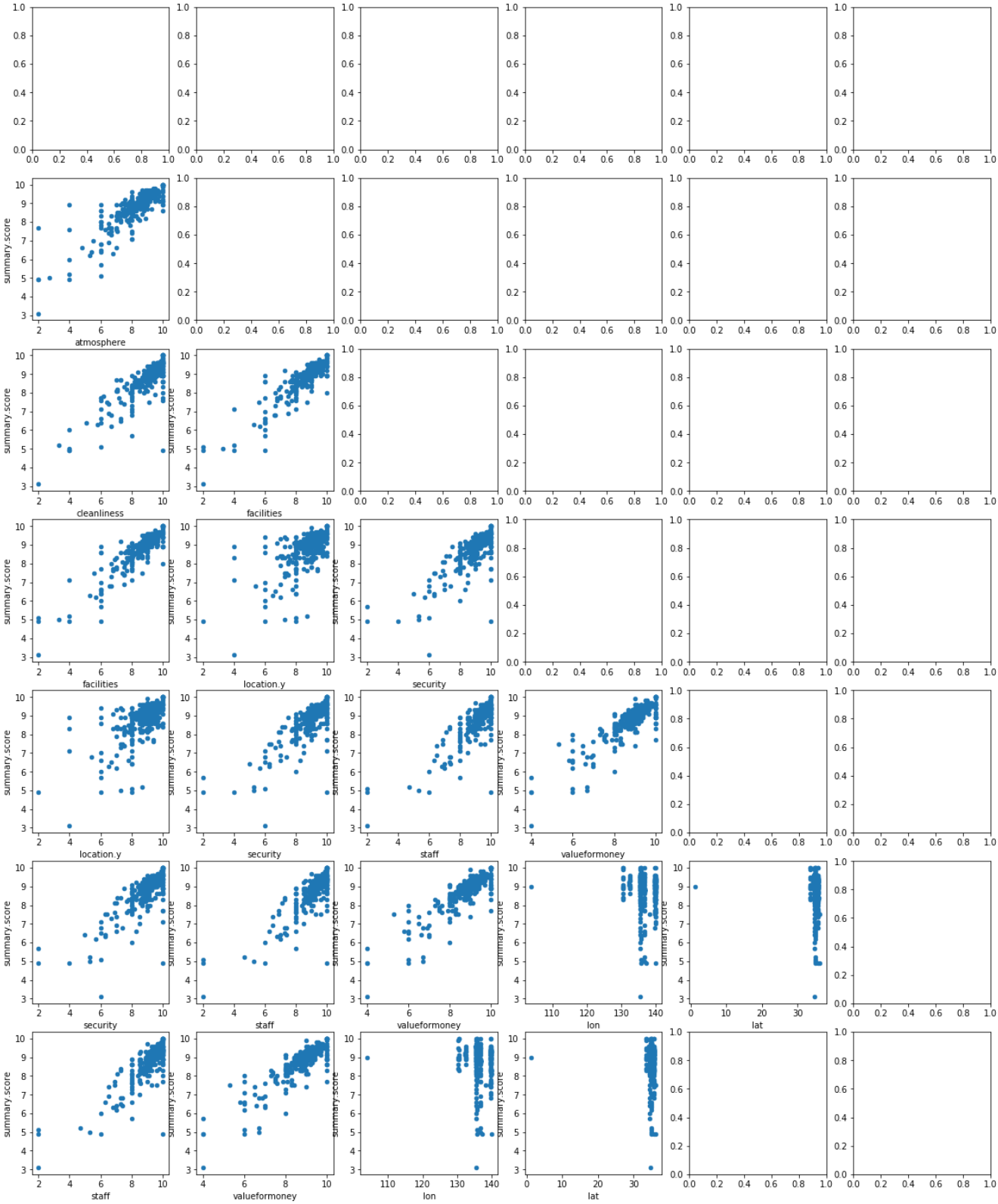
hostel.name, City, rating.band, price.from, summary.score, atmosphere, cleanliness, facilities, location.y, security, staff, valueformoney, lon, lat

Заполнение пропусков

Пропущенные данные заполнялись на основе средних значений, а в случае категориальных признаков – как самый популярный.

Визуализация

Зависимость главного значения от всех числовых



Mushroom Classification

Описание входных данных

- class – съедобный/ядовитый
- cap-shape – форма шляпки
- cap-surface – поверхность шляпки
- cap-color – цвет шляпки
- bruises – повреждения
- odor – запах
- gill-attachment – прикрепление внутренней пластины шляпки
- gill-spacing – интервал на внутренней пластине шляпки
- gill-size – размер внутренней пластины шляпки
- gill-color – цвет внутренней пластины шляпки
- stalk-shape – форма ножки
- stalk-root – основание ножки
- stalk-surface-above-ring – поверхность ножки над кольцом
- stalk-surface-bellow-ring – поверхность ножки под кольцом
- stalk-color-above-ring – цвет ножки над кольцом
- stalk-color-bellow-ring – цвет ножки под кольцом
- veil-type – тип завесы
- veil-color – цвет завесы
- ring-number – число колец
- ring-type – тип кольца
- spore-print-color – цвет спор
- population – популяция
- habitat – ареал

Анализ данных

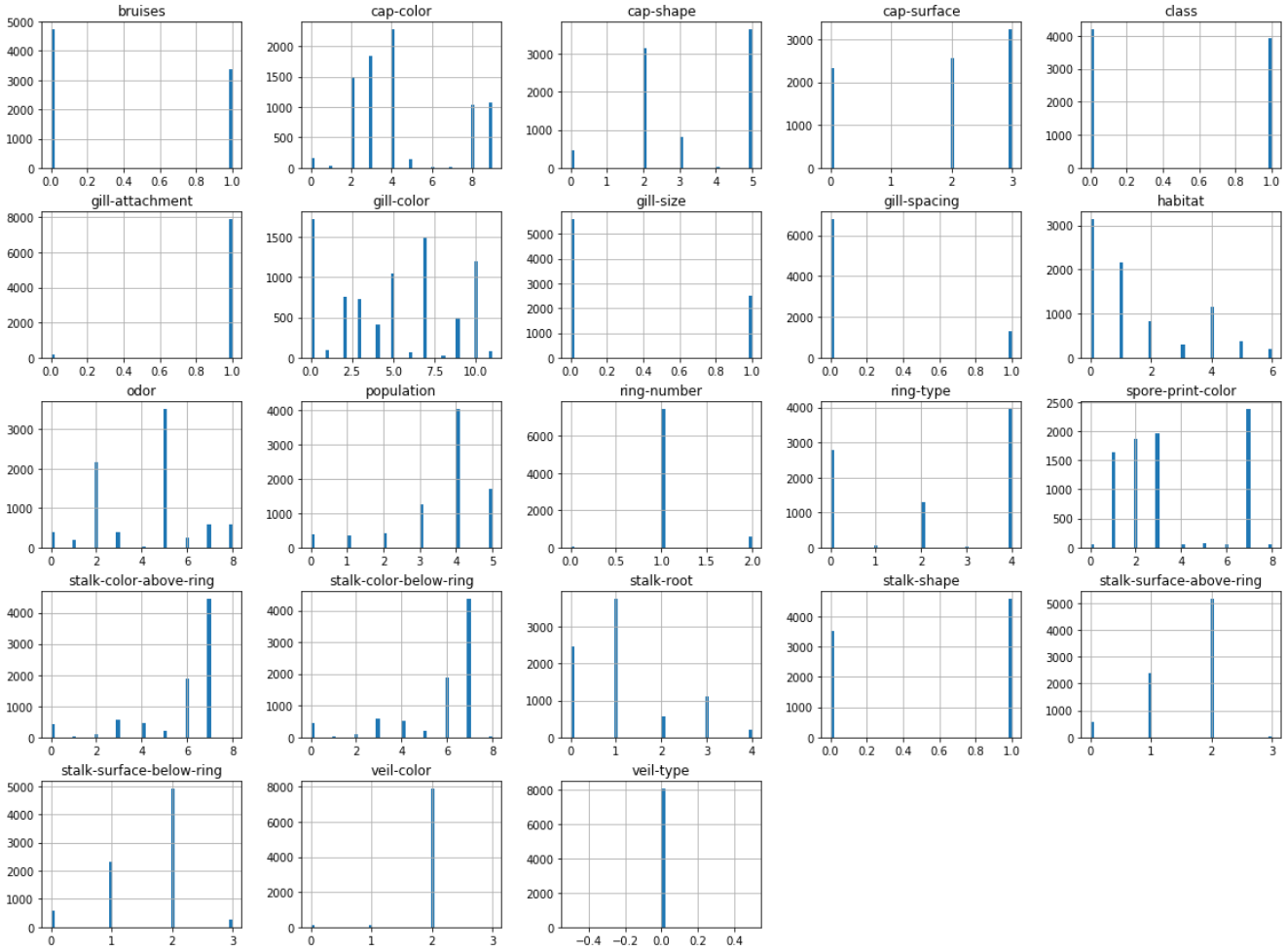
Типы признаков

- Все признаки количественные
- Иследуемое значение: population

Размер

- Строк: 8124
- Столбцов: 23

Распределение признаков с числовыми полями



Решаемая задача

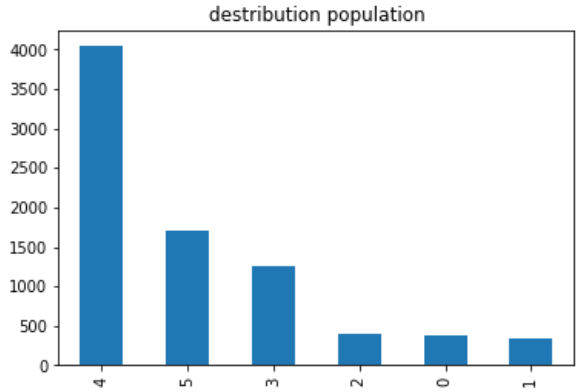
Классифицировать population.

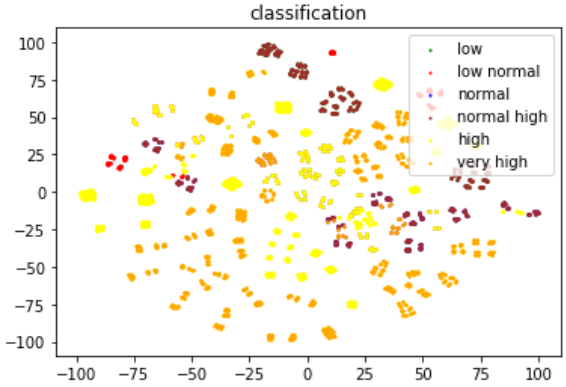
Проблемы

Изначально была проблема с данными: все они были типа object, и мне было проблематично с ними полноценно работать. Изменение типа решило эту проблему.

Визуализация

Распределение по кластерам.





Вывод

В ходе лабораторной работы были проанализированы два датасета, данные для каждого из них были подготовлены для поставленной задачи. Также были показаны распределение признака, который предстоит исследовать, и его зависимость от других признаков.