Московский авиационный институт
(национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 806 «Вычислительная математика и программирование»

Лабораторная работа №0 по курсу «Искусственный интеллект»

Студент: Д.А. Тарпанов
Преподаватели: Д. В. Сошников
С. Х. Ахмед
Группа: М8О-407Б-19
Дата:
Оценка:
Подпись:

Москва, 2022

Лабораторная работа №0

Задача: В данной лабораторной работе вы выступаете в роли предприимчивого начинающего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба. Поэтому тщательно работайте. И главное, день промедления и вас опередит ваш конкурент, да и сплагиаченная работа отразится на репутации По сути в данной лабораторной работе вы выполняете часть работы BI системы.

1 Ход работы

Я выбрал набор данных Water quality classification [1] для выполнения лабораторной работы. Требуется предсказать, является ли вода безопасной, основываясь на её составе.

Признаки в наборе данных:

1. aluminium - опасен, если содержание больше 2.8
2. ammonia - опасен, если содержание больше 32.5
3. arsenic - опасен, если содержание больше 0.01
4. barium - опасен, если содержание больше 2
5. cadmium - опасен, если содержание больше 0.005
6. chloramine - опасен, если содержание больше 4
7. chromium - опасен, если содержание больше 0.1
8. copper - опасен, если содержание больше 1.3
9. flouride - опасен, если содержание больше 1.5
10. bacteria - опасен, если содержание больше 0
11. viruses - опасен, если содержание больше 0
12. lead - опасен, если содержание больше 0.015
13. nitrates - опасен, если содержание больше 10
14. nitrites - опасен, если содержание больше 1
15. mercury - опасен, если содержание больше 0.002
16. perchlorate - опасен, если содержание больше 56
17. radium - опасен, если содержание больше 5
18. selenium - опасен, если содержание больше 0.5
19. silver - опасен, если содержание больше 0.1
20. uranium - опасен, если содержание больше 0.3

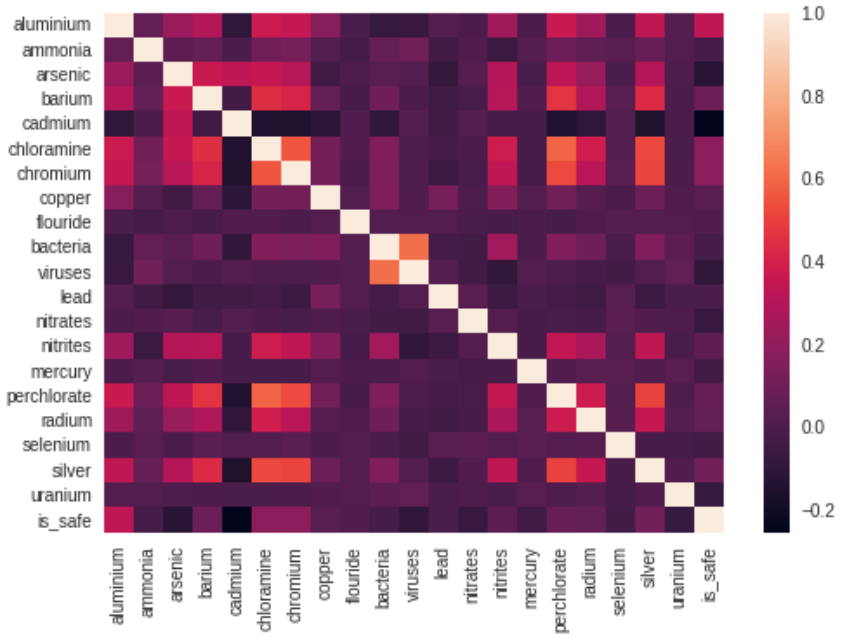
21. is_safe - класс воды {0 - не безопасна, 1 - безопасна}

Перед выявлением зависимостей между признаками проверяю целостность набора данных:

```
RangeIndex: 7999 entries,0 to 7998
Data columns (total 21 columns):
#   Column       Non-Null Count  Dtype
--- ------      -------------- -----
0   aluminium    7999 non-null   float64
1   ammonia      7999 non-null   object
2   arsenic      7999 non-null   float64
3   barium       7999 non-null   float64
4   cadmium      7999 non-null   float64
5   chloramine   7999 non-null   float64
6   chromium     7999 non-null   float64
7   copper       7999 non-null   float64
8   flouride     7999 non-null   float64
9   bacteria     7999 non-null   float64
10  viruses      7999 non-null   float64
11  lead         7999 non-null   float64
12  nitrates     7999 non-null   float64
13  nitrites     7999 non-null   float64
14  mercury      7999 non-null   float64
15  perchlorate  7999 non-null   float64
16  radium       7999 non-null   float64
17  selenium     7999 non-null   float64
18  silver       7999 non-null   float64
19  uranium      7999 non-null   float64
20  is_safe      7999 non-null   object
dtypes: float64(19),object(2)
memory usage: 1.3+ MB
```

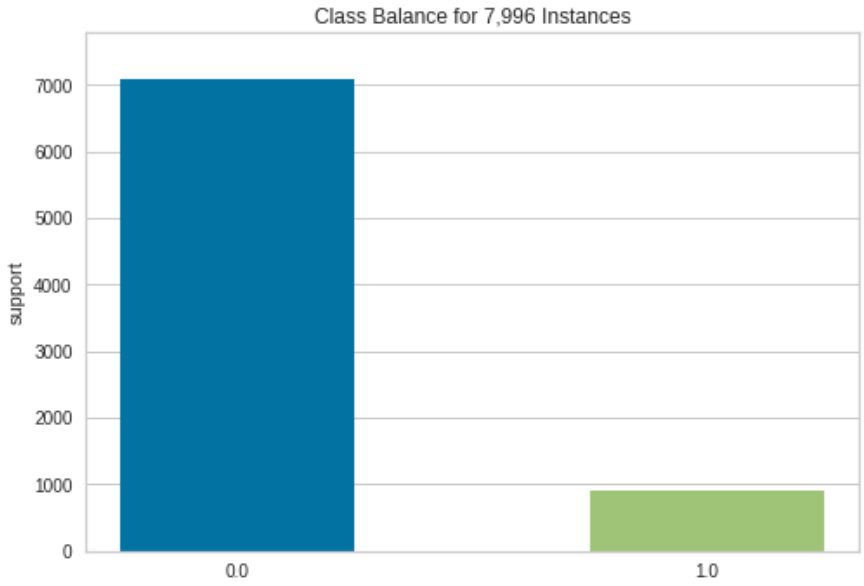
В наборе есть неполные данные, в пропусках записана строка NUM!, их необходимо удалить.

Построю корреляционную матрицу для признаков:

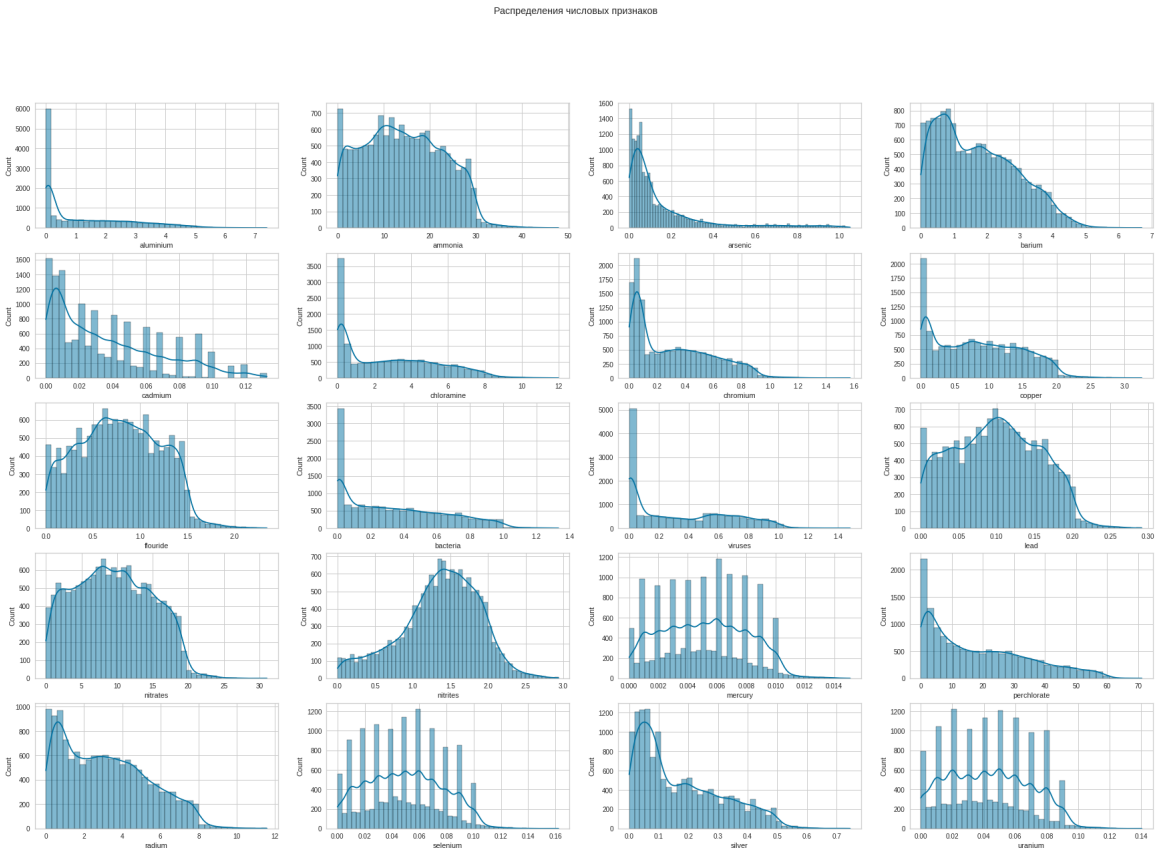


Так же построю гистограммы для числовых признаков:

Видно, что много признаков распределены плохо. Соотношение классов:

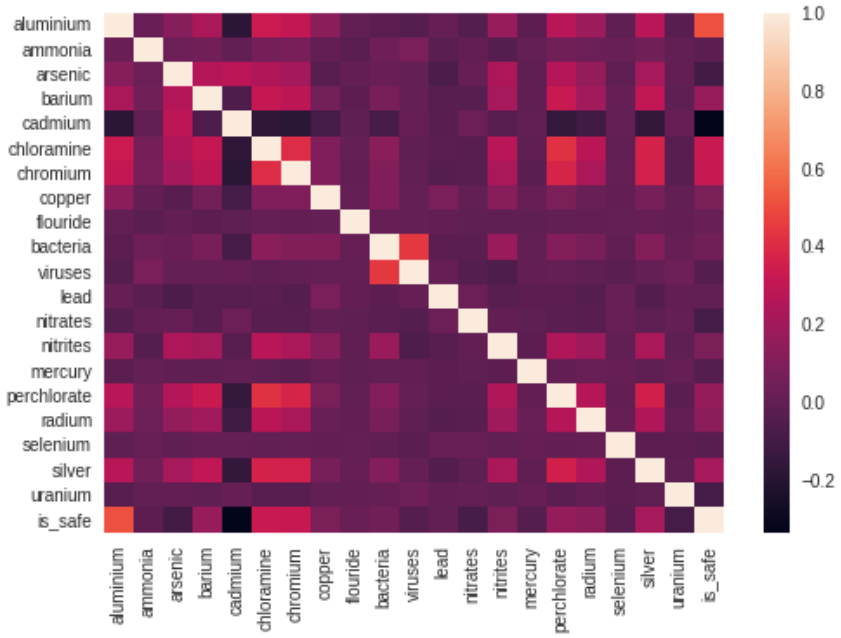


Классы несбалансированы, нужно провести аугментацию. Для этого посчитаем матожидание и дисперсию каждого признака, потом сгенерируем нормально распределенные новые признаки по заданным матожиданию и дисперсии. Новые распределения:

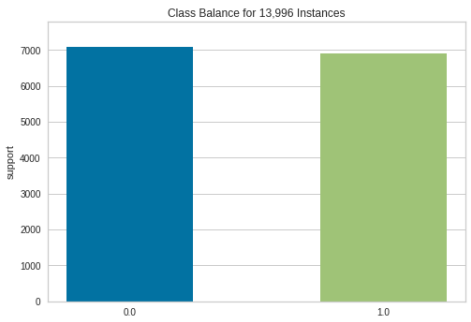


Видно, что распределения сильно лучше не стали, однако, забегая вперед, скажу, что такая аугментация повысила recall с 30 процентов до 80-90.

Корреляционная матрица почти не изменилась:



Баланс классов:



Классы стали сбалансированными

2 Выводы

В ходе выполнения лабораторной работы я освежил в памяти курс математической статистики: гистограмму, корелляцию и корреляционную матрицу для наборов данных. Так же я изучил библиотеку Pandas, она оказалась очень удобной для анализа данных.

Набор данных оказался не самым лучшим, с такими распределениями признаков получить высокую точность у линейных моделей может быть проблематично.

Был проанализирован набор данных Water quality [1], результаты получились закономерные: безопасность воды равномерно скореллированна с другими признаками, но нашлись интересные зависимости: количество бактерий и количество вирусов, содержание хрома и хлорамина, или перхлората и серебра. Наверное, это можно описать химическими реакциями между веществами.

Список литературы

[1] *Water quality | Kaggle*
URL: `https://www.kaggle.com/datasets/mssmartypants/water-quality`
(дата обращения: 10.11.2022).

[2] *Exploratory data analysis with Pandas — mlcourse.ai*
URL: `https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html`
(дата обращения: 10.11.2022).