

Министерство науки высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»
(Университет ИТМО)

Факультет технологий искусственного интеллекта

Лабораторная работа №4
Анализ данных на предмет выбросов и дрифтов

Выполнили:

Гончаренко Данила Олегович, группа J4150
Стрельницкая Татьяна Викторовна, группа J4140

Преподаватель:

Старобыховская Анастасия Александровна

Санкт-Петербург
2024

Оглавление

Задание	2
Основные этапы.....	3
Вывод.....	8

Задание

В рамках данной лабораторной работы поставлена цель – проанализировать датасет предыдущих лабораторных работ на предмет шумов, отклонений и дрейфов.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Взять задачу из прошлых лабораторных и датасет,
2. Проанализировать данные для датасета на наличие шумов, выбросов/аномалий, дрейфов.
3. Выбрать алгоритм определения дрейфа и применить его,
- 3+. Прикрутить инструмент мониторинга.
4. Проанализировать полученные результаты. Сделать выводы и описать почему они могли получиться именно такими.
6. Написать отчёт.

Основные этапы

Была взята задача классификации качества вина.

Входные переменные - различные физико-химические свойства вина – представлены в таблице 1.

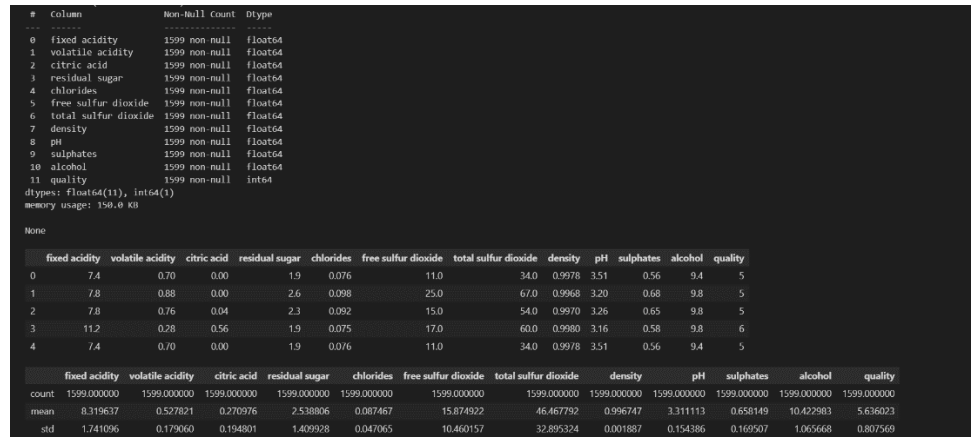
Таблица 1 – Набор входных данных для модели логистической регрессии

Название характеристики	Описание
Fixed Acidity	Уровень кислотности, который остается после ферментации
Volatile Acidity	Количество уксусной кислоты, влияющей на вкус
Citric Acid	Лимонная кислота – усиливает вкус, придает свежесть
Residual Sugar	Сахар, остающийся после ферментации
Chlorides	Хлориды – содержание соли
Free Sulfur Dioxide	Свободный диоксид серы – действует как противомикробное средство
Total Sulfur Dioxide	Общее количество диоксида серы
Density	Плотность вина, связанная с содержанием алкоголя и сахара
pH	Уровень кислотности
Sulphates	Сульфаты, способствующие микробной стабильности
Alcohol	Процент алкоголя
Quality	Сенсорная оценка качества (от 0 до 10)

Выходная переменная: показатель качества (0-10), определяемый на основе сенсорных данных.

Задача – определить выбросы и отклонения в датасете.

Взят датасет качества вина из прошлых лабораторных работ.



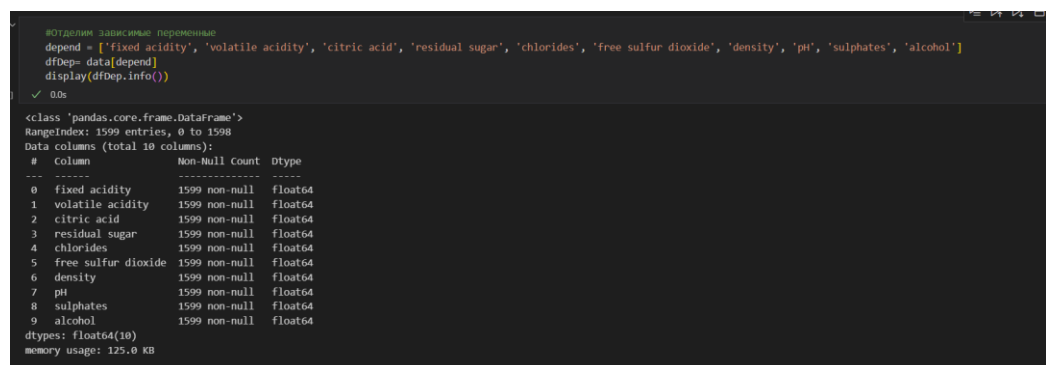
```
# column      Non-Null Count  Dtype
---  -
0 fixed acidity    1599 non-null  float64
1 volatile acidity  1599 non-null  float64
2 citric acid      1599 non-null  float64
3 residual sugar   1599 non-null  float64
4 chlorides        1599 non-null  float64
5 free sulfur dioxide 1599 non-null  float64
6 total sulfur dioxide 1599 non-null  float64
7 density          1599 non-null  float64
8 pH              1599 non-null  float64
9 sulphates        1599 non-null  float64
10 alcohol         1599 non-null  float64
11 quality         1599 non-null  int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538006	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.402928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569

Рисунок 1 – Исходный датасет качества вина.

Далее отделяем зависимые переменные от целевой переменной. Для анализа были выделены следующие переменные: 'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol'.



```
#отделим зависимые переменные
depend = ['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol']
dfDep= data[depend]
display(dfDep.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 10 columns):
#   column      Non-Null Count  Dtype
---  -
0 fixed acidity    1599 non-null  float64
1 volatile acidity  1599 non-null  float64
2 citric acid      1599 non-null  float64
3 residual sugar   1599 non-null  float64
4 chlorides        1599 non-null  float64
5 free sulfur dioxide 1599 non-null  float64
6 density          1599 non-null  float64
7 pH              1599 non-null  float64
8 sulphates        1599 non-null  float64
9 alcohol         1599 non-null  float64
dtypes: float64(10)
memory usage: 125.0 KB
```

Рисунок 2 – Подготовленный датасет зависимых переменных

Проводим анализ выбросов при помощи ящиков с усами. При помощи ящиков с усами были выявлены выбросы по каждому показателю.

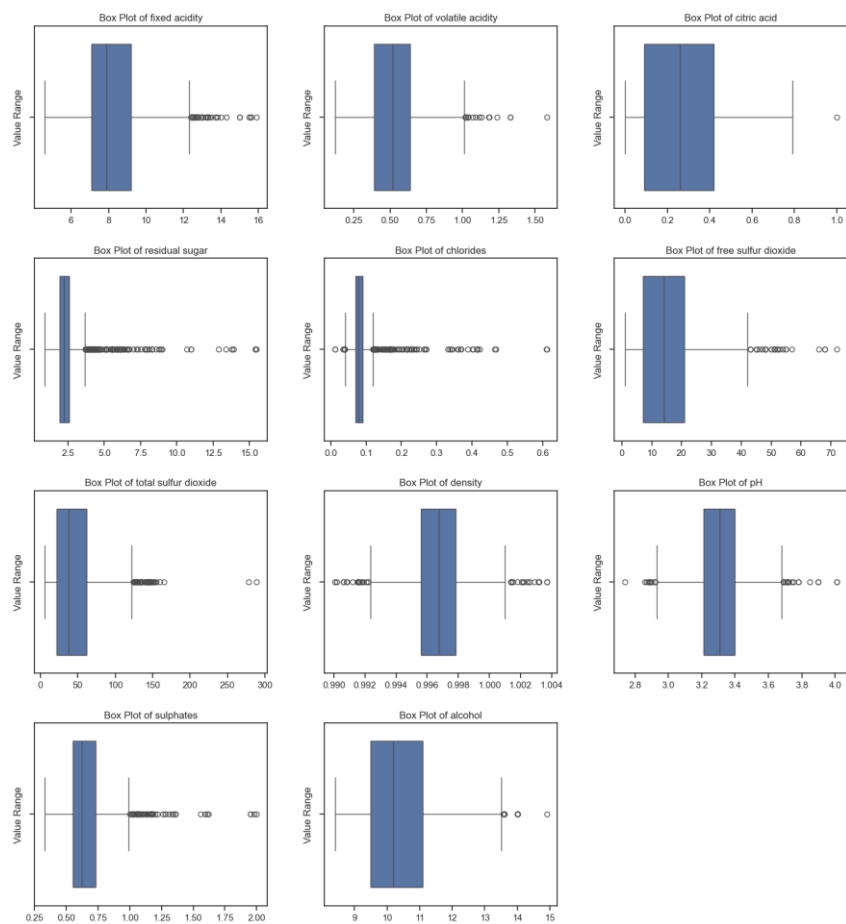


Рисунок 3 – Ящики с усами по зависимым переменным.

Затем для нахождения дрифта датасет был разделен на два тестовых датасата..

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
799	9.4	0.500	0.34	3.6	0.082	5.0	14.0	0.99870	3.29	0.52	10.7
800	7.2	0.610	0.08	4.0	0.082	26.0	108.0	0.99641	3.25	0.51	9.4
801	8.6	0.550	0.09	3.3	0.068	8.0	17.0	0.99735	3.23	0.44	10.0
802	5.1	0.585	0.00	1.7	0.044	14.0	86.0	0.99264	3.56	0.94	12.9
803	7.7	0.560	0.08	2.5	0.114	14.0	46.0	0.99710	3.24	0.66	9.6

Рисунок 4 – Тестовые датасеты

Для нахождения дрифта данных был выбран алгоритм Колмагорова-Смирнова.

	Feature	KS-statistic	P-value
7	density	0.464770	9.666870e-79
10	alcohol	0.310853	1.057344e-34
0	fixed acidity	0.274646	3.738210e-27
4	chlorides	0.201912	1.064279e-14
2	citric acid	0.194121	1.231822e-13
3	residual sugar	0.174332	4.762430e-11
8	pH	0.152927	1.326065e-08
6	total sulfur dioxide	0.105842	2.186888e-04
5	free sulfur dioxide	0.094387	1.486029e-03
1	volatile acidity	0.067300	4.881952e-02
9	sulphates	0.066383	5.574066e-02

Рисунок 5 – Результаты работы алгоритма Колмагорова-Смирнова.

Для мониторинга был использован EVENDITLY AI. Для анализа дрифтов датасет был разделен на два равных тестовых сабсета. В результате анализа было найдено 7 колонок, имеющих дрифты: chlorides, alcohol, citric acid, pH, free sulfur dioxide, density.

Dataset Drift		
Dataset Drift is detected. Dataset drift detection threshold is 0.5		
11 Columns	7 Drifted Columns	0.636 Share of Drifted Columns
Data Drift Summary		
Drift is detected for 63.636% of columns (7 out of 11).		

Рисунок 6 – Результаты анализа датасета на наличие дрифтов.




>	chlorides	num			Detected	K-S p_value	0.034077
>	alcohol	num			Detected	K-S p_value	0.023518
>	citric acid	num			Detected	K-S p_value	0.002481
>	pH	num			Detected	K-S p_value	0.002249
>	free sulfur dioxide	num			Detected	K-S p_value	0.000118

Рисунок 7 – Результаты анализа датасета по переменным.

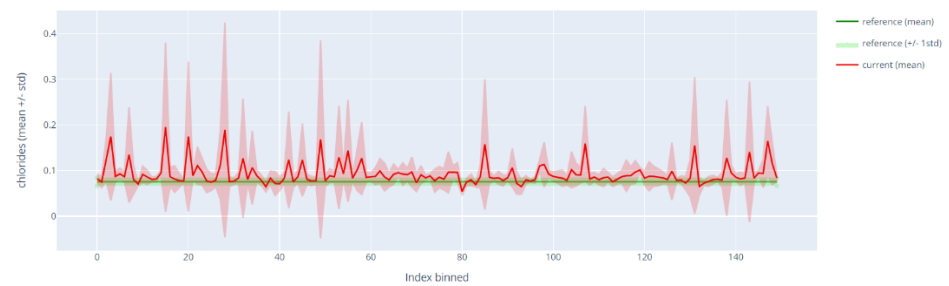


Рисунок 8 – Data drift chlorides.

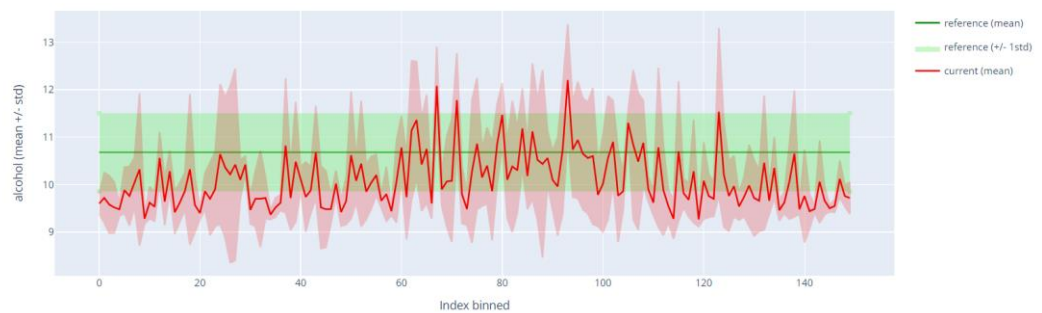


Рисунок 9 – Data drift alcohol

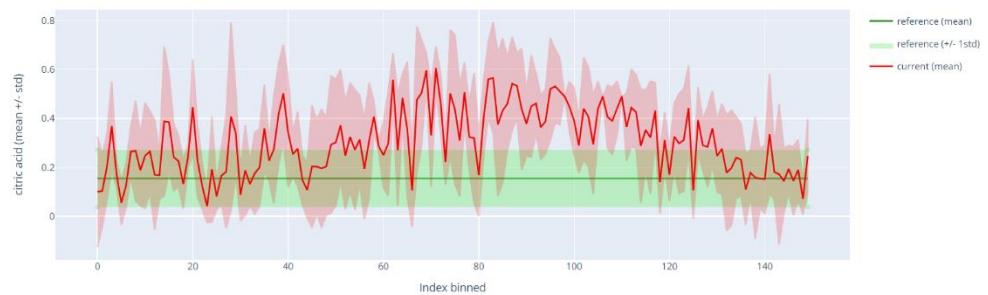


Рисунок 10 – Data drift citric acid

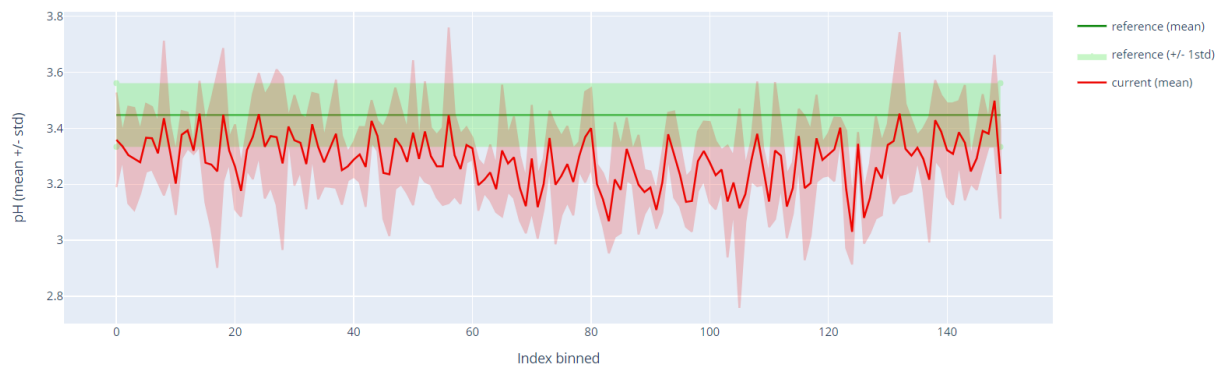


Рисунок 11 – Data drift pH

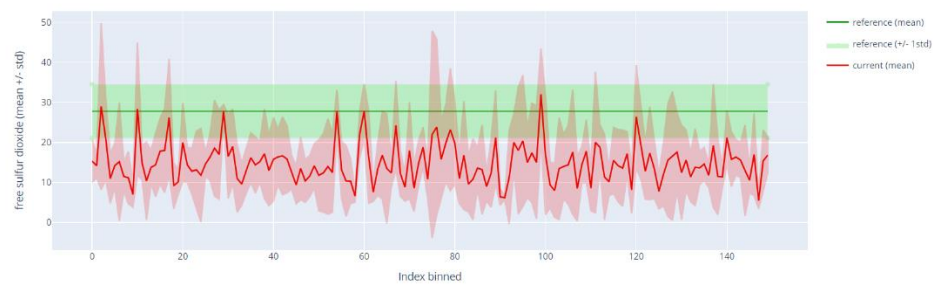


Рисунок 12 – Data drift free sulfur dioxide

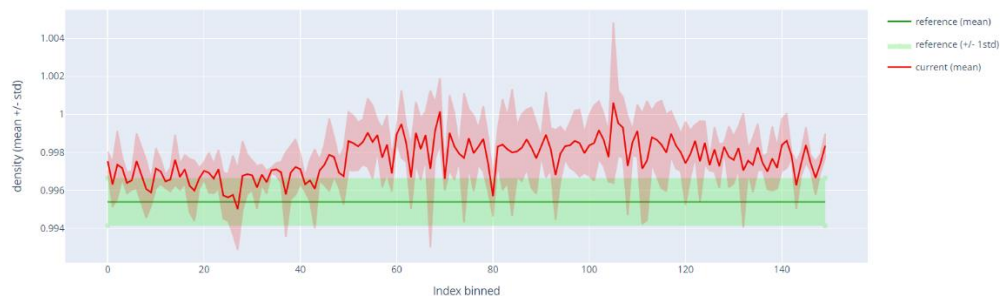


Рисунок 13 – Data drift density

Выводы

В результате определения дрейфа при помощи алгоритма Колмагорова-Сморнова можно сделать вывод, что данные имеют не выбросы каждому показателю, а дрейфы наблюдаются в 7 показателях. [Ссылка](#)