**Московский государственный технический**

**университет им. Н.Э. Баумана**

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по рубежному контролю №1

«Технологии разведочного анализа и обработки данных»

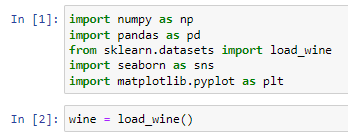
Вариант №2

|  |  |  |
| --- | --- | --- |
| Выполнил: |  | Проверил: |
| студент группы ИУ5-61Б |  | преподаватель каф. ИУ5 |
| Бондаренко Денис Константинович |  | Гапанюк Юрий Евгеньевич |
|  |  |  |
| Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Дата:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  | Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_  Дата:\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  |  |  |
|  |  |  |

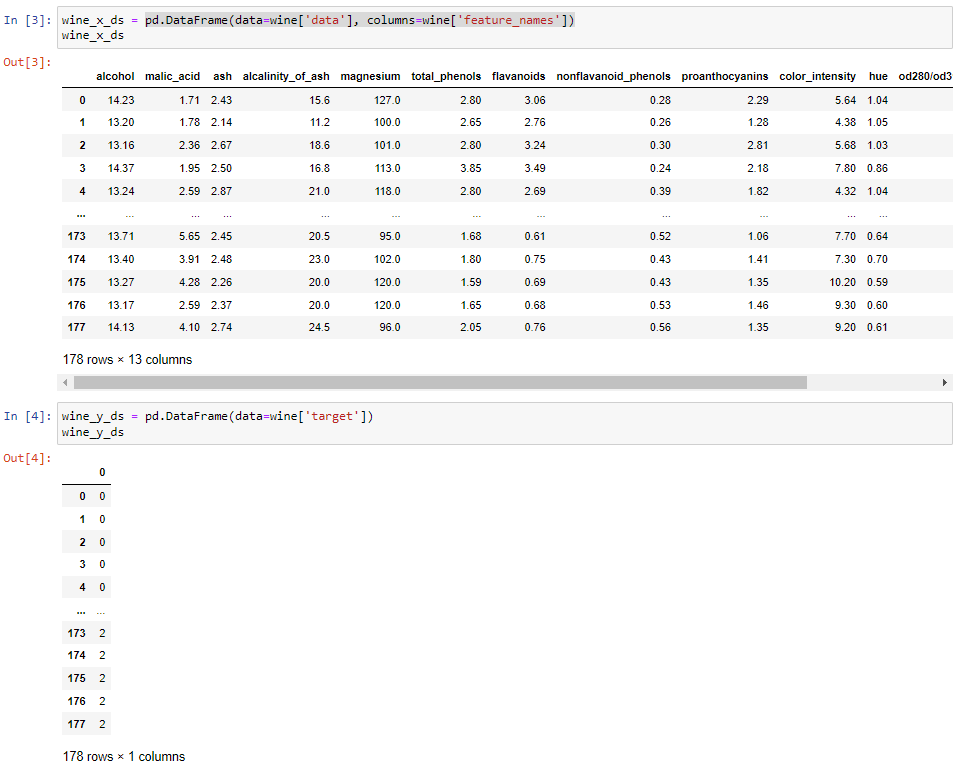
Москва, 2023 г.

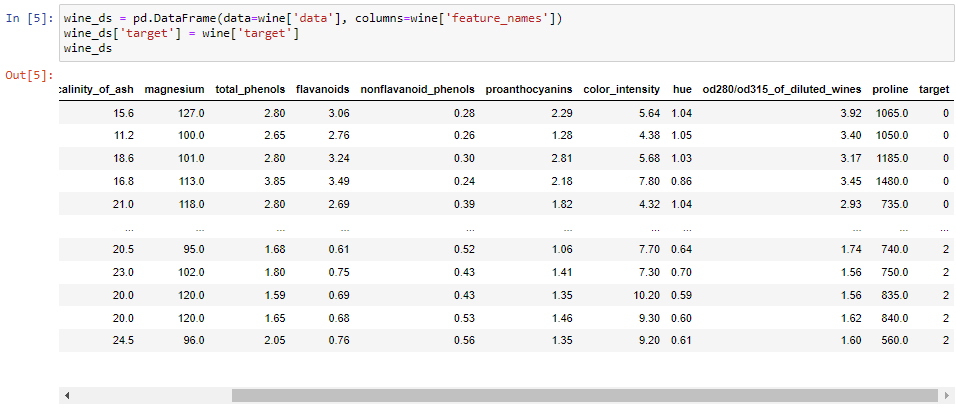
**Выполнение работы**

Для выполнения задачи проведения корреляционного анализа данных был представлен набор данных sklearn wine dataset, загруженный с помощью функции load\_wine().

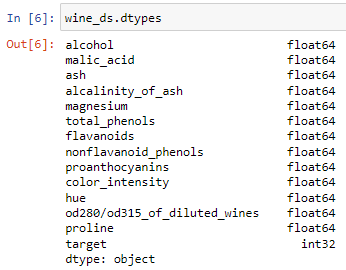


Был создан датафрейм, содержащий 13 нецелевых признаков и 1 целевой – класс вина.

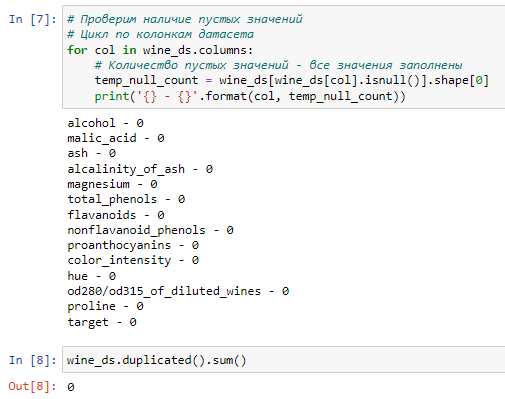




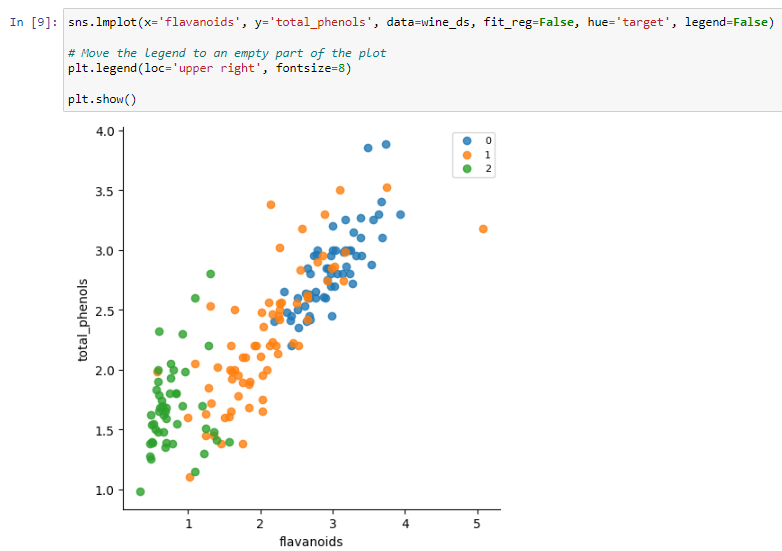
Типы данных всех полей являются числовыми.



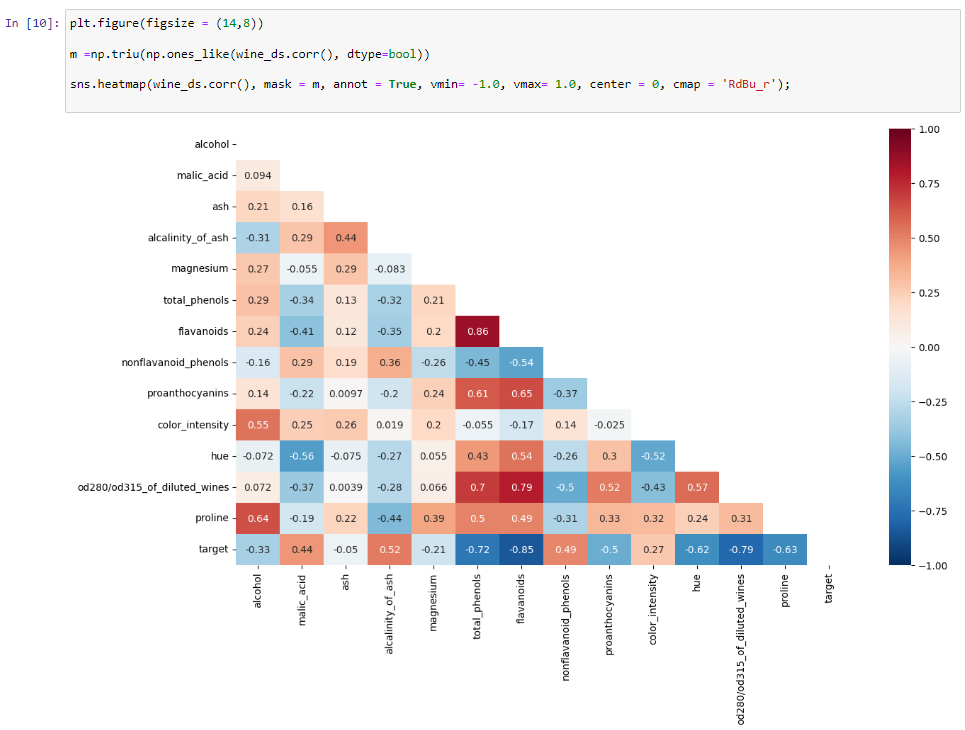
В наборе данных отсутствуют пропуски и дубликаты.



Для колонок flavanoids и total\_phenols был построен график “Диаграмма рассеяния”.



Для визуализации корреляционной матрицы была использована “тепловая карта”.



С целевым признаком наиболее сильную корреляцию имеют признаки “flavanoids” (-0,85), “od280/od315\_of\_diluted\_wines” (-0,79), “total\_phenols” (-0,72), “proline” (-0,63) и “hue” (-0,62). Эти признаки будут наиболее информативными при построении моделей машинного обучения. Целевой признак отчасти коррелирует с признаками “alcalinity\_of\_ash” (0,52), “proanthocyanins” (-0,5), “nonflavanoid\_fenols” (0,49) и “malic\_acid” (0,44). Эти признаки также стоит использовать при обучении модели. Признаки “alcohol” (-0,33), “color\_intensity” (0,27), “magnesium” (-0,21) и “ash” (-0,05) слабо коррелируют с целевым признаком и могут негативно сказаться на модели машинного обучения, поэтому, скорее всего, их стоит исключить из модели.

Но не все признаки, которые имеют сильную и среднюю корреляцию с целевым признаком, стоит использовать для построения модели машинного обучения. Между признаками “flavanoids” и “total\_phenols” наблюдается очень сильная корреляция (0,86). Это связано с тем, что флавоноиды относятся к классу полифенолов. Поэтому из этих двух признаков стоит оставить тот, который имеет наибольшую корреляцию с целевым признаком, т.е. “flavanoids”. Остальные нецелевые признаки не коррелируют друг с другом так сильно и между ними не наблюдается почти линейной зависимости.

Таким образом, на основе признаков “flavanoids”, “od280/od315\_of\_diluted\_wines”, “proline”, “hue”, “alcalinity\_of\_ash”, “proanthocyanins”, “nonflavanoid\_phenols” и “malic\_acid” могут быть построены модели машинного обучения, первые четыре признака могут иметь наиболее весомый вклад в их обучение.