Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по рубежному контролю №1

Тема: «Технологии разведочного анализа и обработки данных.»

*Вариант 3*

Студент:

Белкина Е.В.

Группа ИУ5-61Б

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 г.

**Задание**

Задача №1.

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Набор данных:

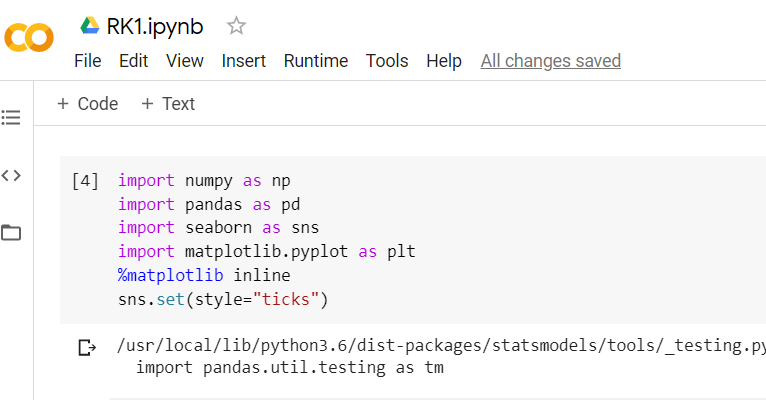
<https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine>

Дополнительные требования по группам:

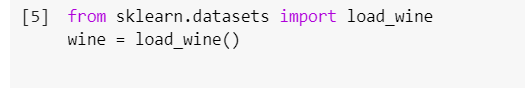
Для студентов групп ИУ5-61Б, ИУ5Ц-81Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

**Выполнение задания**

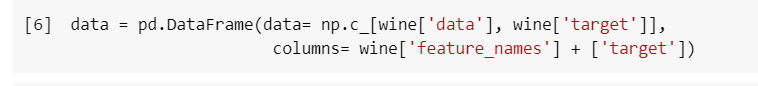
1. Импортируем необходимые библиотеки с помощью команды import.



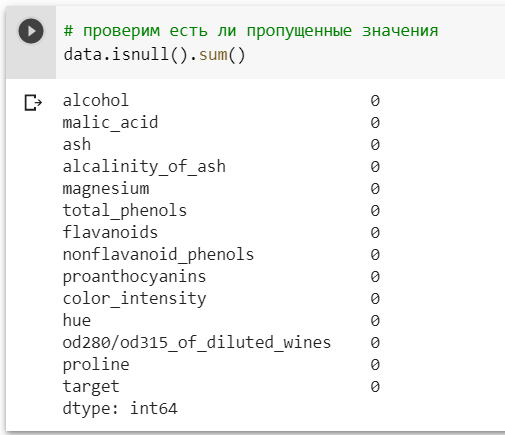
1. Импортируем датасет load\_wine из sklearn в соответствии с заданием варианта



1. Преобразуем датасет Scikit-learn в Pandas Dataframe



1. Проверим наличие пропусков данных

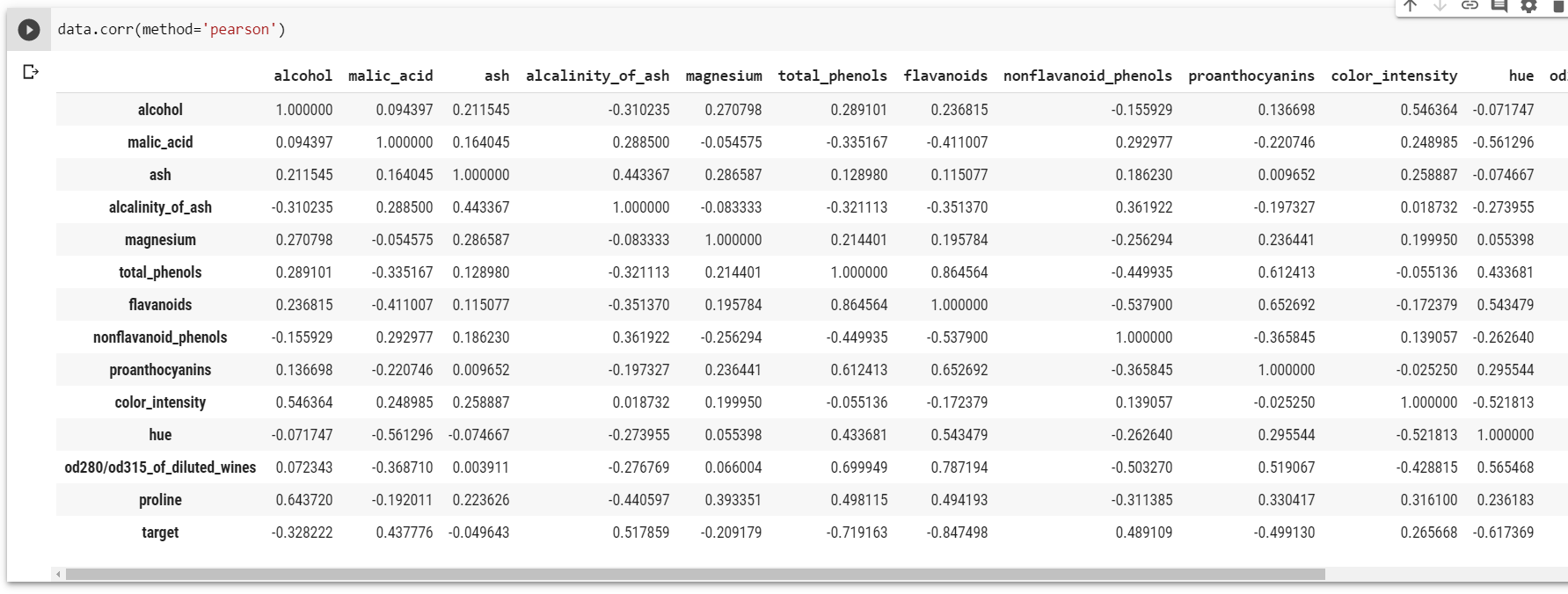


Можем видеть, что пропуски данных в датасете отсутствуют.

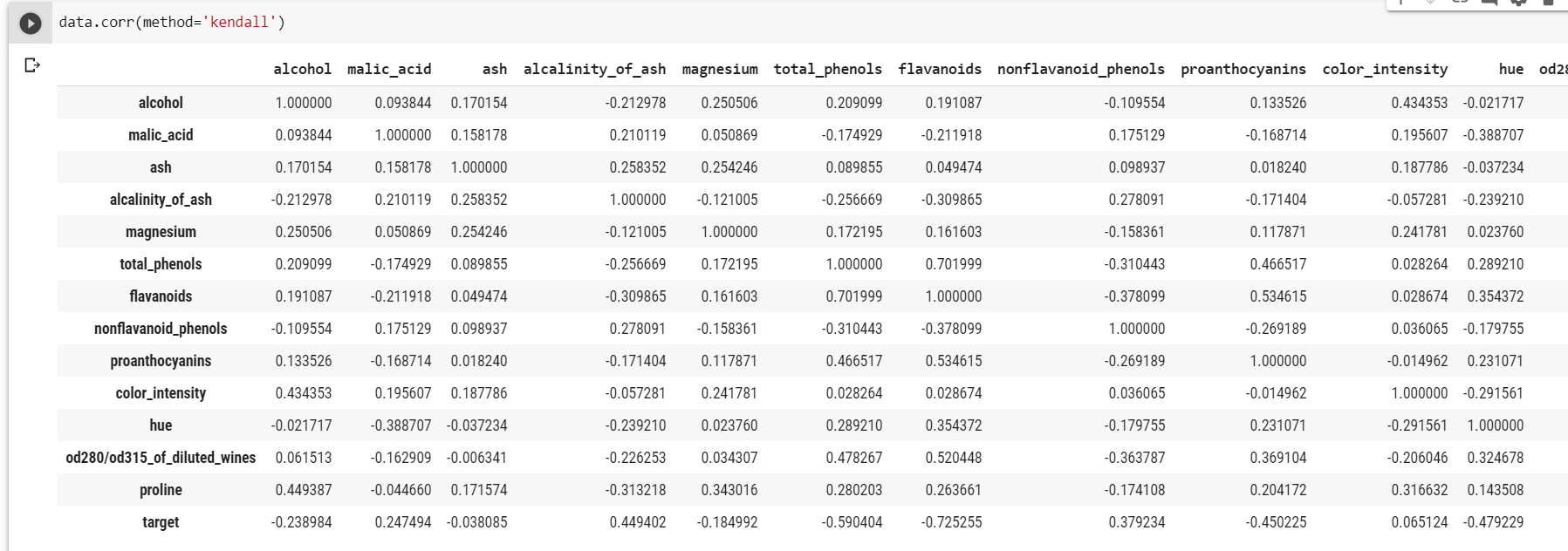
Таким образом, мы можем построить корректную корреляционную матрицу.

1. Проведём корреляционный анализ

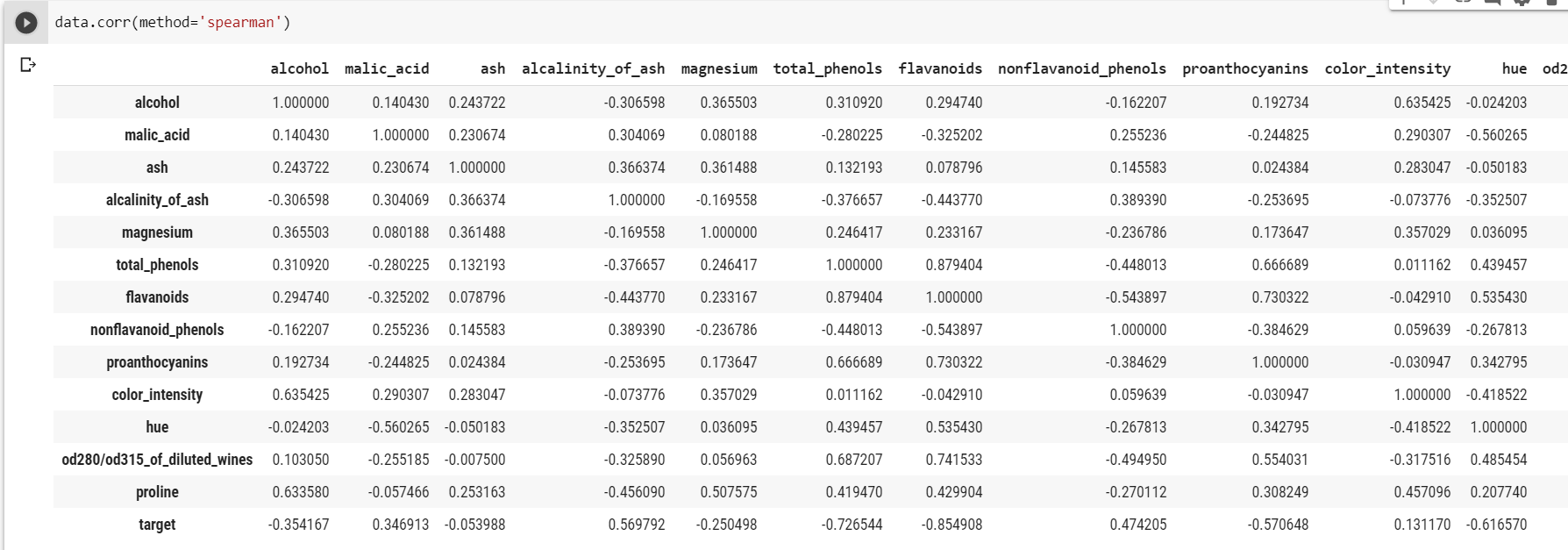
Матрица с коэффициентом корреляции Пирсона:

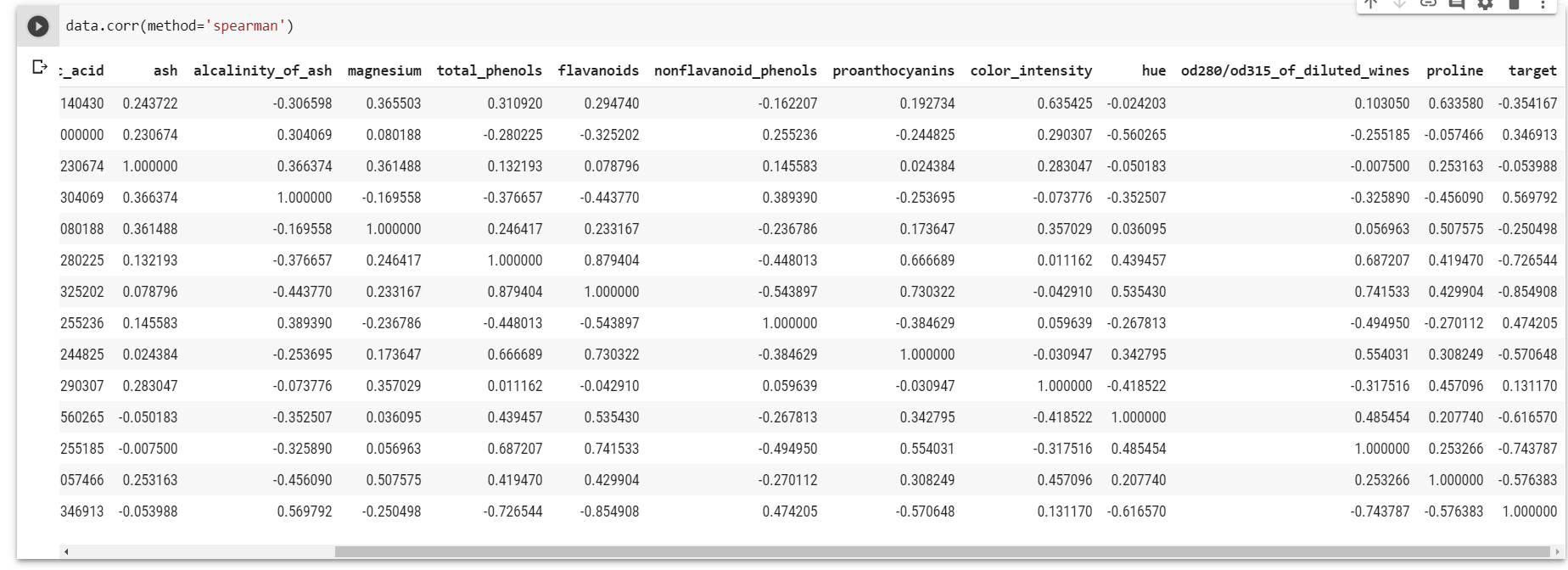


Матрица с коэффициентом корреляции Кендалла:

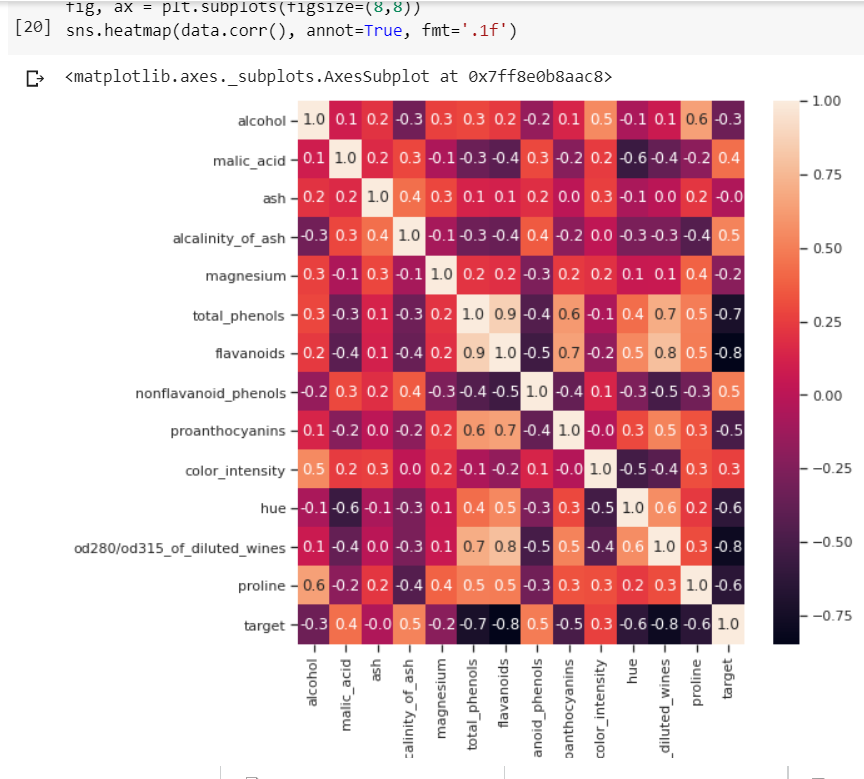


Матрица с коэффициентом корреляции Спирмана:





Тепловая карта корреляционной матрицы:



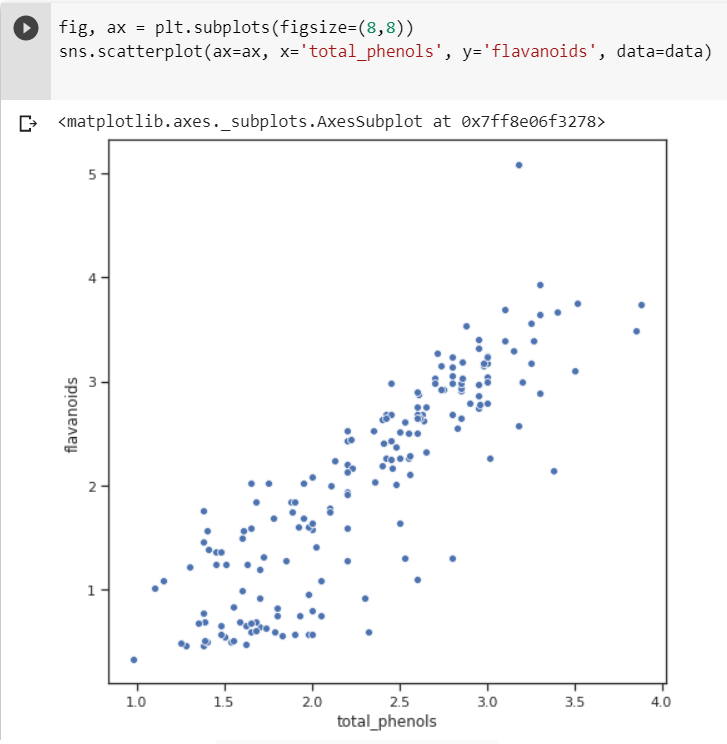
**Корреляционный анализ:**

Необходимо понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (колонка "target"). Именно эти признаки будут наиболее информативными для моделей машинного обучения. Значительнее большинства признаков с target коррелируют alcanlinity\_of\_ash (0.5), anoid\_phenols (0.5), malic\_acid (0.4). Эти признаки следует оставить в модели.

Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. В данном датасете такими являются flavanolds (-0.8), total\_phenols (-0.7), od280/od315\_of\_diluted\_wines (-0.8), hue (-0.6), proline (-0.6).

Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если признаки линейно зависимы, то для построения модели из них выбирают какой-то один признак. В нашем наборе данных максимально коррелируют между собой flavonoids и total\_phenols (0.9), они имеют практически линейную зависимость. Поэтому для построения модели лучше оставить только один из этих признаков, наиболее коррелирующий с целевым. Но оба эти признака так плохо коррелируют с target, что имеет смысл убрать каждый из них.

Диаграмма рассеяния для колонок total\_phenols и flavonoids:



1. Выводы

На данном наборе данных (датасет Wine recognition dataset из sklearn) возможно удачно построить модель машинного обучения, так как он не имеет пропусков данных, содержит признаки, значительно коррелирующие с целевым и не зависимые линейно. Для улучшения построенной модели имеется возможность отказаться от линейно зависимых и слабо коррелирующих с целевым признаков.