Московский государственный технический университет им. Н.Э. Баумана Кафедра

«Системы обработки информации и управления»



Домашняя работа по дисциплине

«Методы машинного обучения»

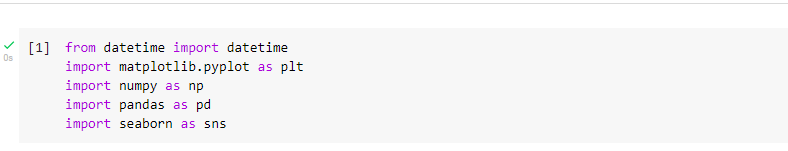
Выполнил:  
студент группы ИУ5-24М  
Зо Хтет Аунг

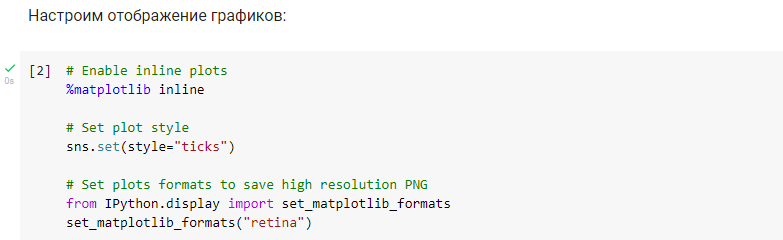
Москва — 2022г.

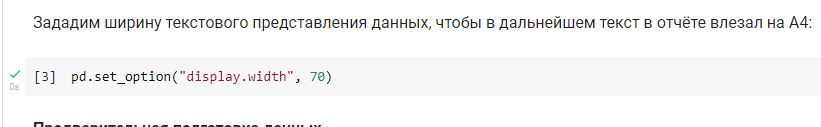
1. **Задание**  
   Требуется выполнить следующие действия [1]:  
   1. Поиск и выбор набора данных для построения моделей машинного обучения.  
   На основе выбранного набора данных студент должен построить модели машинного обучения для решения или задачи классификации, или задачи регрессии.  
   2. Проведение разведочного анализа данных. Построение графиков, необходимых  
   для понимания структуры данных. Анализ и заполнение пропусков в данных.  
   3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных  
   признаков, улучшающих качество моделей.  
   4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. В зависимости  
   от набора данных, порядок выполнения пунктов 2, 3, 4 может быть изменен.  
   5. Выбор метрик для последующей оценки качества моделей. Необходимо выбрать  
   не менее двух метрик и обосновать выбор.  
   6. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. Необходимо использовать не менее трех моделей, хотя бы одна из которых  
   должна быть ансамблевой.  
   7. Формирование обучающей и тестовой выборок на основе исходного набора данных.  
   8. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров. Производится обучение моделей на основе обучающей выборки и  
   оценка качества моделей на основе тестовой выборки.  
   9. Подбор гиперпараметров для выбранных моделей. Рекомендуется подбирать  
   не более 1-2 гиперпараметров. Рекомендуется использовать методы кроссвалидации. В зависимости от используемой библиотеки можно применять функцию GridSearchCV, использовать перебор параметров в цикле, или использовать другие методы.

10. Повторение пункта 8 для найденных оптимальных значений гиперпараметров.  
Сравнение качества полученных моделей с качеством baseline-моделей.  
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

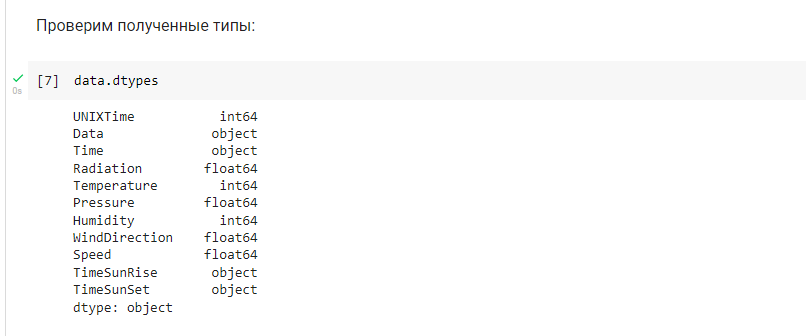
1. **Ход выполнения работы**  
   **2.1. Выбор набора данных**  
   В качестве набора данных используются метрологические данные с метеостанции  
   HI-SEAS (Hawaii Space Exploration Analog and Simulation) за четыре месяца (с сентября  
   по декабрь 2016 года) и использовался в соревновании Space Apps Moscow 2017 в категории «You are my Sunshine» для построения приложения для предсказания мощности  
   солнечного излучения и планирования работы исследовательской станции [2, 3]. Данный набор данных доступен по следующему адресу: https://www.kaggle.com/dronio/  
   SolarEnergy.
   * 1. **Текстовое описание набора данных**  
        Выбранный набор данных состоит из одного файла SolarPrediction.csv, содержащего все данные датасета. Данный файл содержит следующие колонки:  
        2  
        • UNIXTime — временная метка измерения в формате UNIX;  
        • Data — дата измерения;  
        • Time — время измерения (в местной временной зоне);  
        • Radiation — солнечное излучение (Вт/м2);  
        • Temperature — температура (°F);  
        • Pressure — атмосферное давление (дюймов ртутного столба);  
        • Humidity — относительная влажность (%);  
        • WindDirection(Degrees) — направление ветра (°);  
        • Speed — скорость ветра (миль/ч);  
        • TimeSunRise — время восхода (в местной временной зоне);  
        • TimeSunSet
     2. **Постановка задачи и предварительный анализ набора данных**  
        Очевидно, что данный набор данных предполагает задачу регрессии, а именно предсказание колонки Radiation — мощности солнечного излучения. При этом:  
        • Колонка UNIXTime сама по себе довольно бесполезна, так как просто монотонно  
        растёт с течением времени, не давая какую-либо информацию для модели машинного обучения. Вместе с тем, колонка Time может быть довольно интересной,  
        особенно вместе с колонками TimeSunRise и TimeSunSet, так как вместе они показывают положение солнца на небе и точно задают возможный максимум солнечной  
        энергии.  
        • Колонка Data могла бы быть полезна, если бы данные были за больший промежуток времени (например, несколько лет), и отражала бы сезонность солнечного  
        излучения. К сожалению в нашем случае она практически полностью бесполезна.
   1. **Проведение разведочного анализа данных**  
      Подключим все необходимые библиотеки:

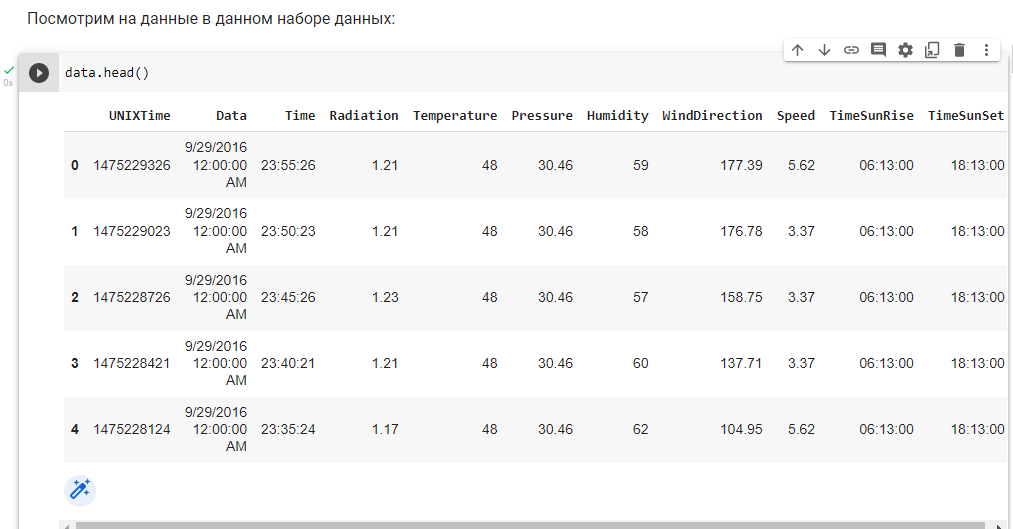


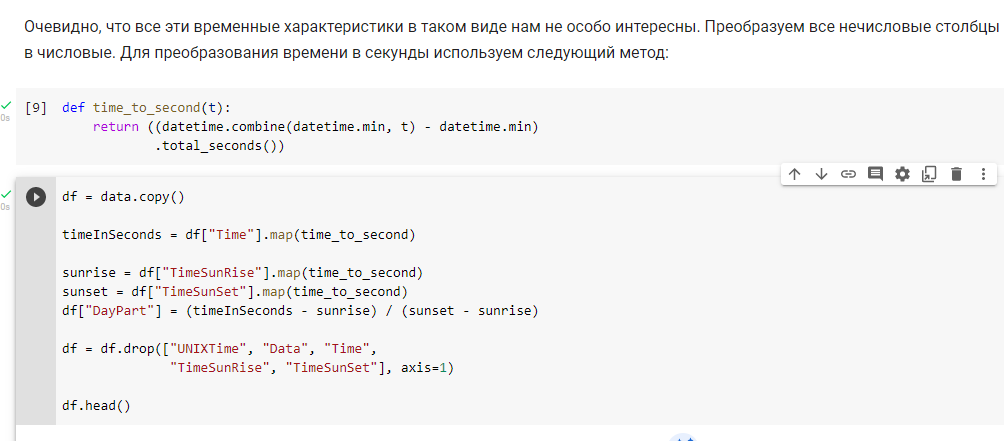


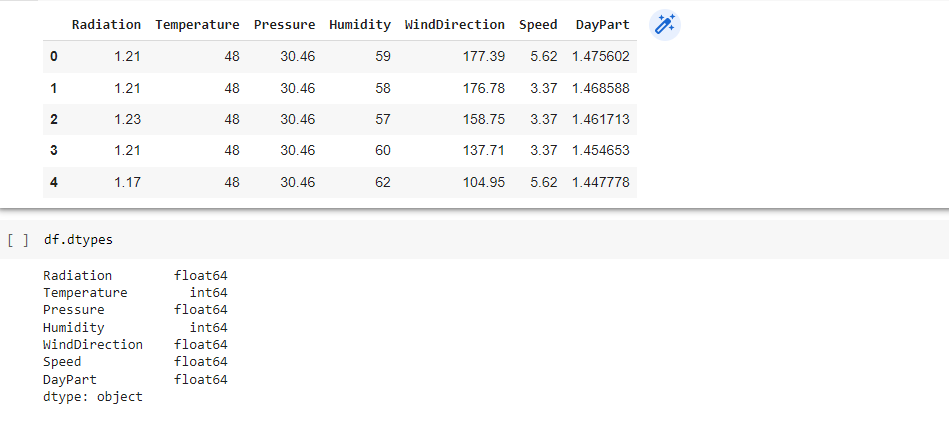


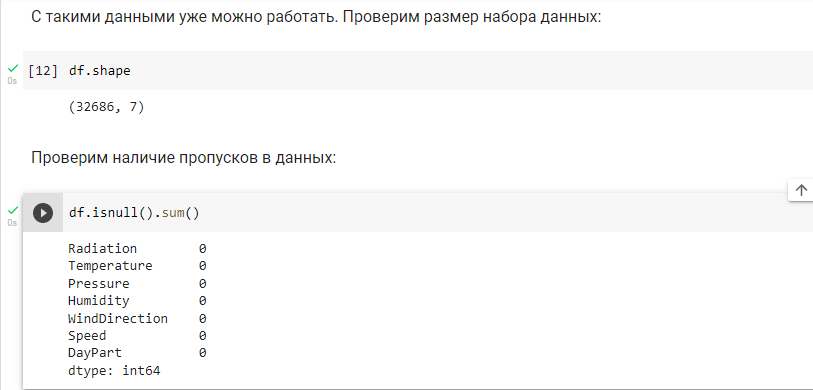
* + 1. **Предварительная подготовка данных**  
       



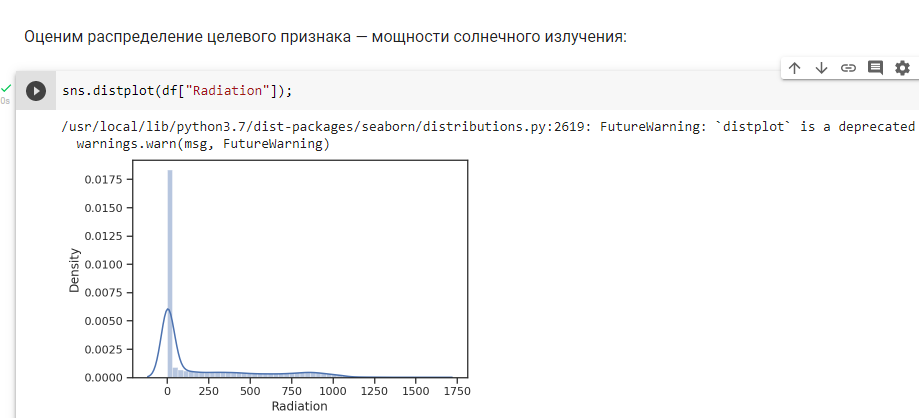




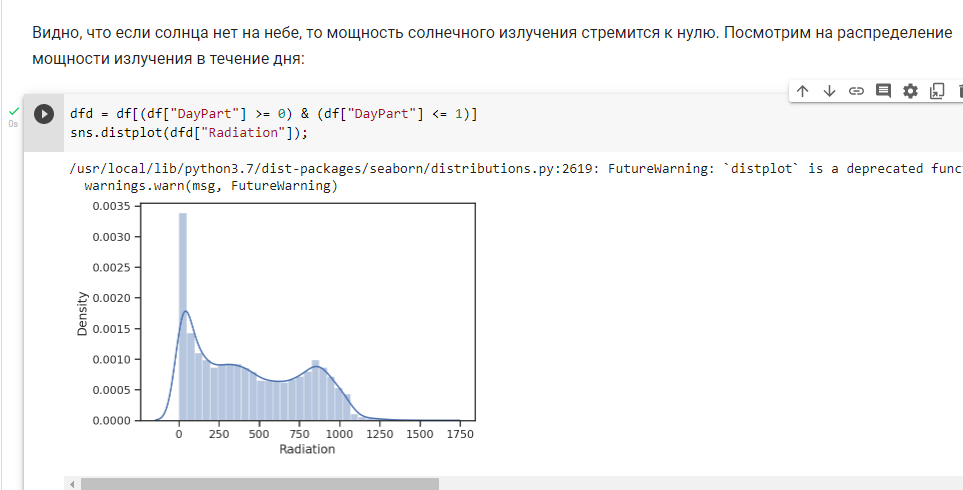


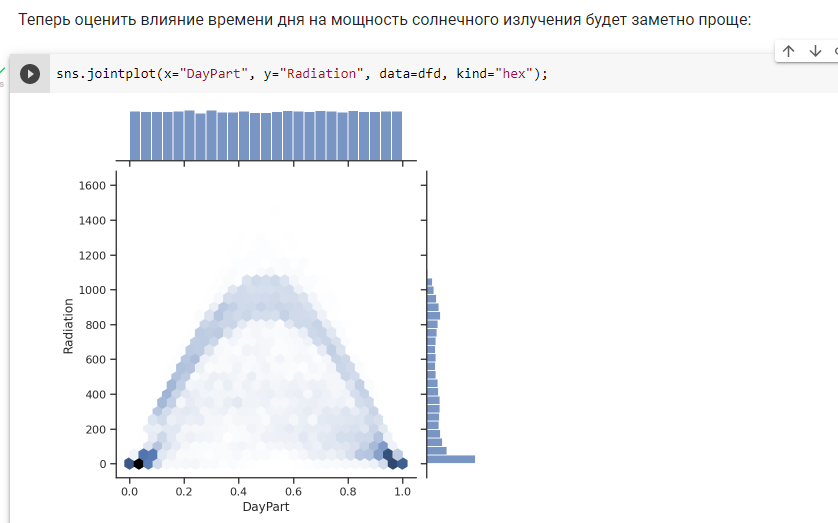


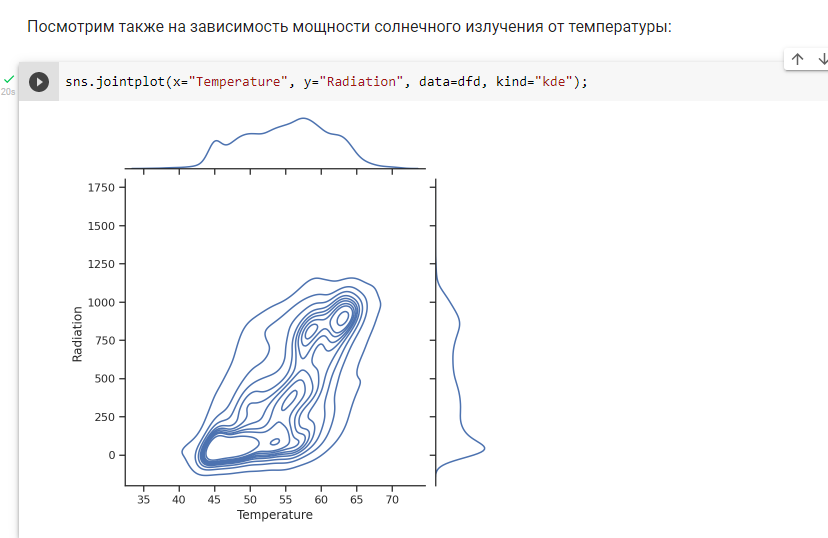
* + 1. **Визуальное исследование датасета**

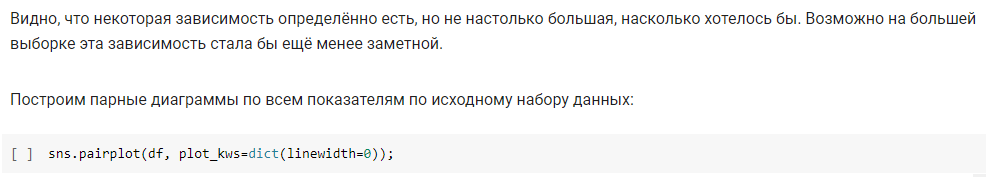


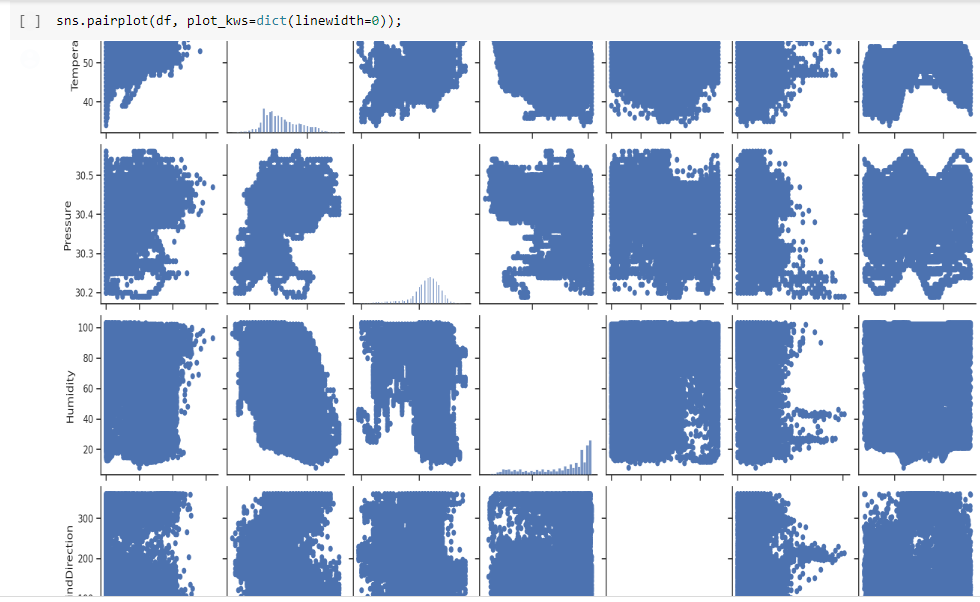






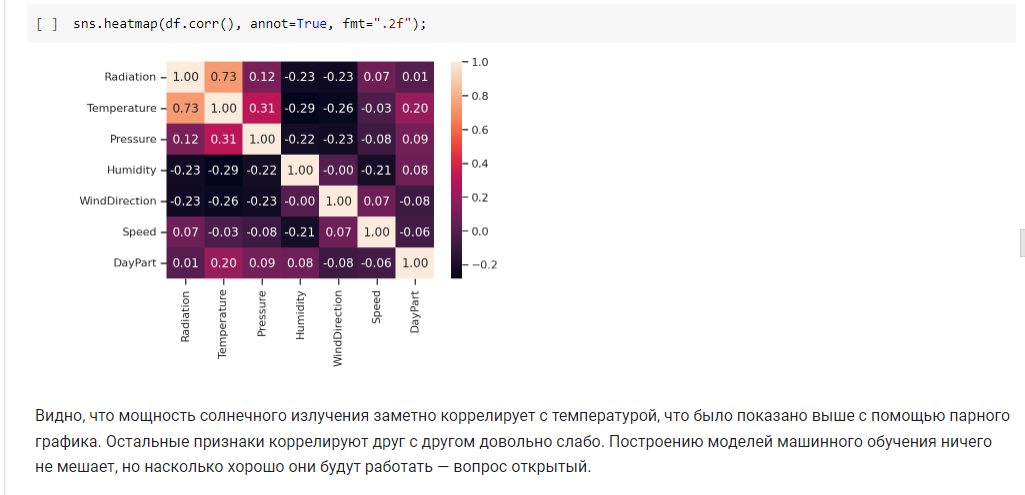




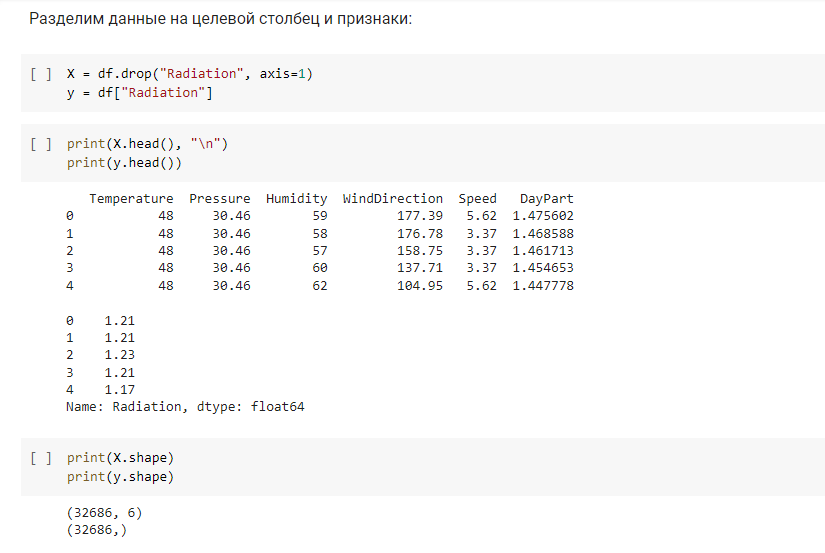
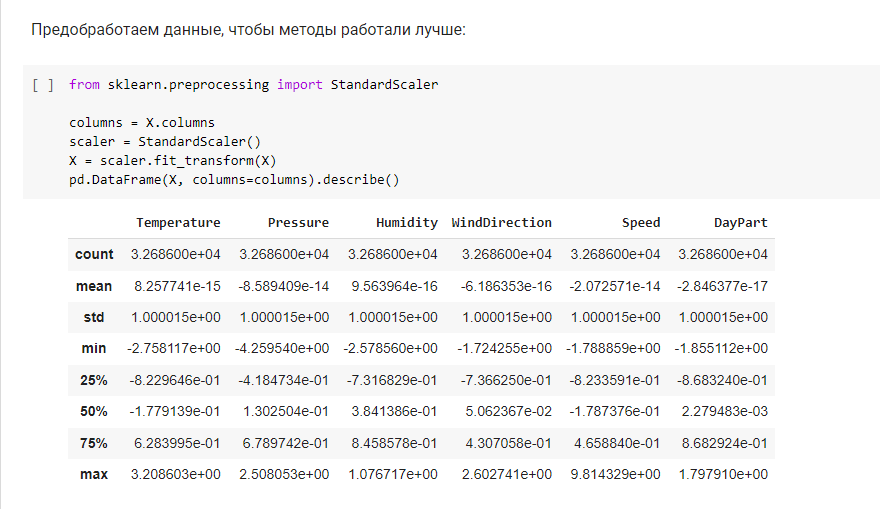


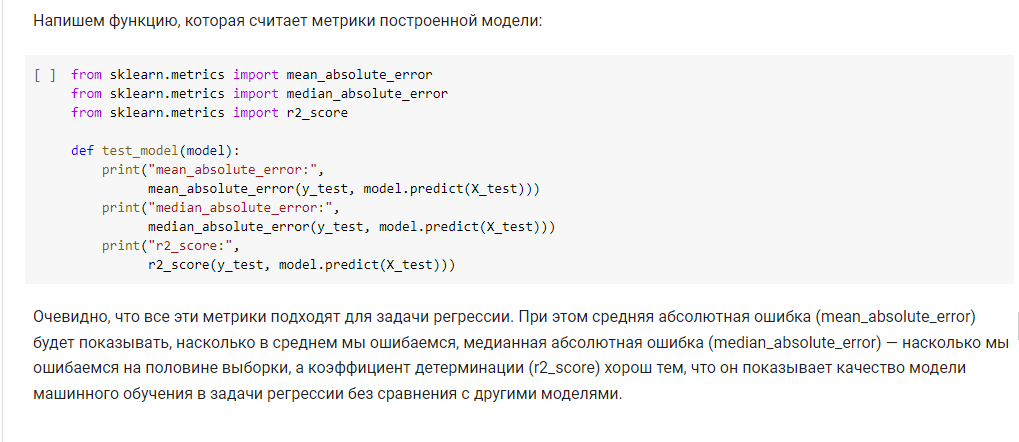
Видно, что зависимости между колонками весьма сложные и в большинстве своём  
нелинейные. Какого-то показателя, точно определяющего мощность излучения, не наблюдается. Вместе с тем чётко видно, что время суток ограничивает мощность излучения сверху, что вполне может быть полезно для модели машинного обучения.

* + 1. **Корреляционный анализ**

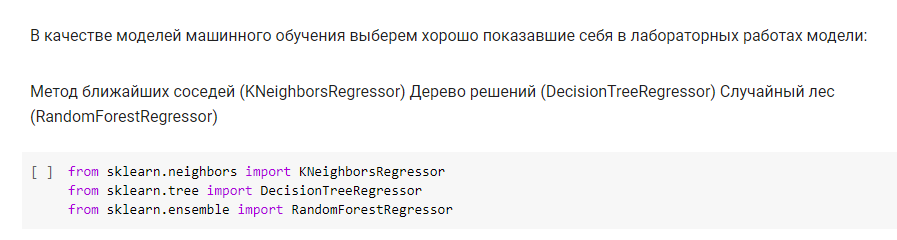
  


* 1. **Подготовка данных для обучения моделей**

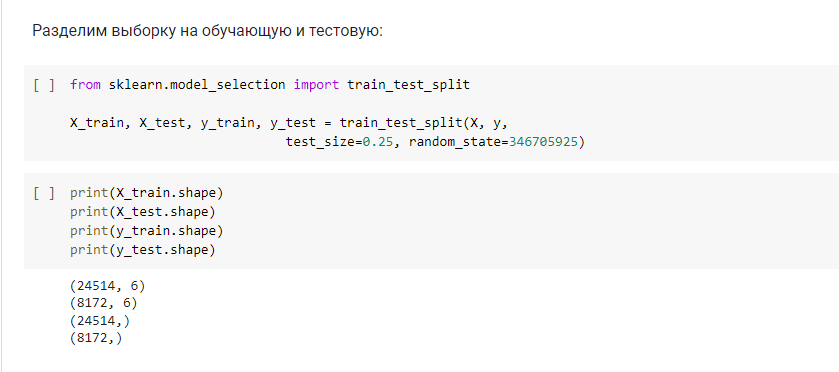
  
  
**2.4. Выбор метрик**



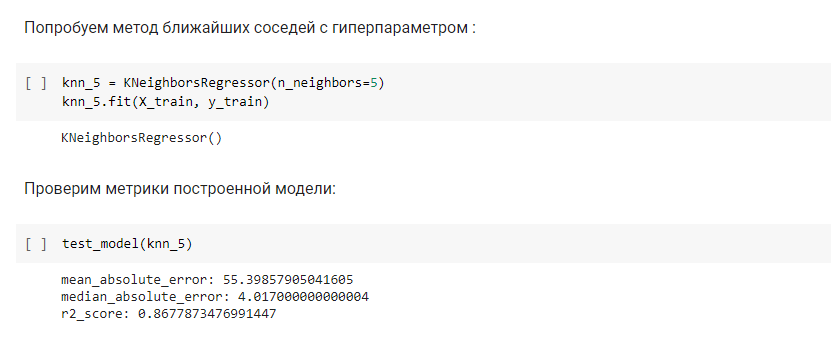
**2.5. Выбор моделей**

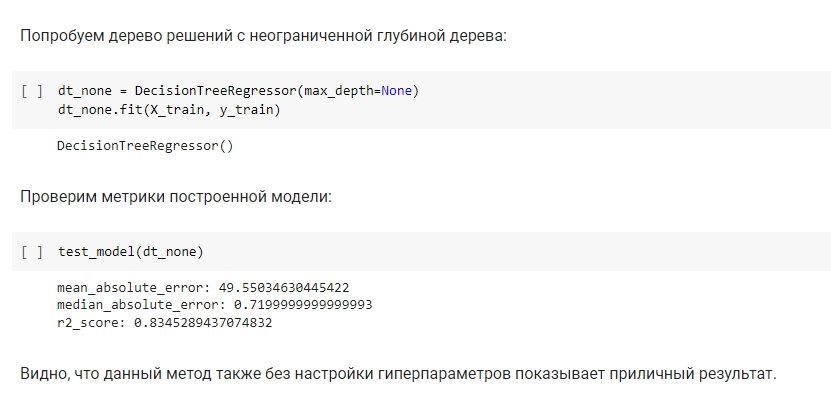


**2.6. Формирование обучающей и тестовой выборок**

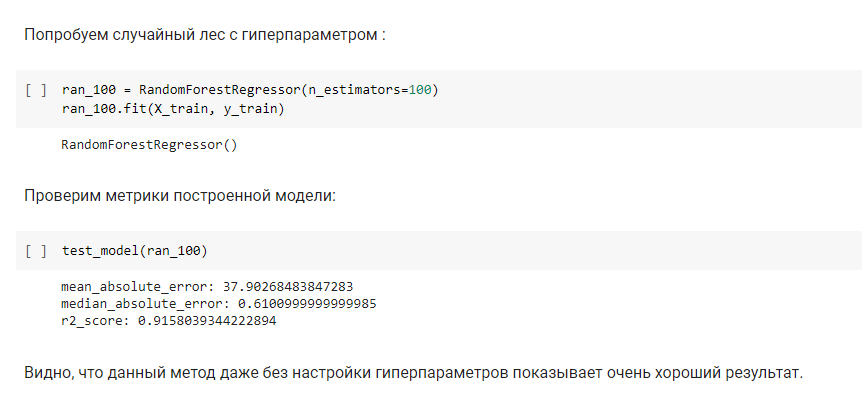


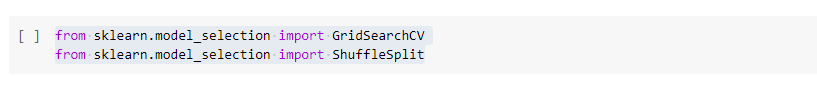
**2.7. Построение базового решения**  
**2.8. Метод** *k* **ближайших соседей**

  
**2.9. Дерево решений**

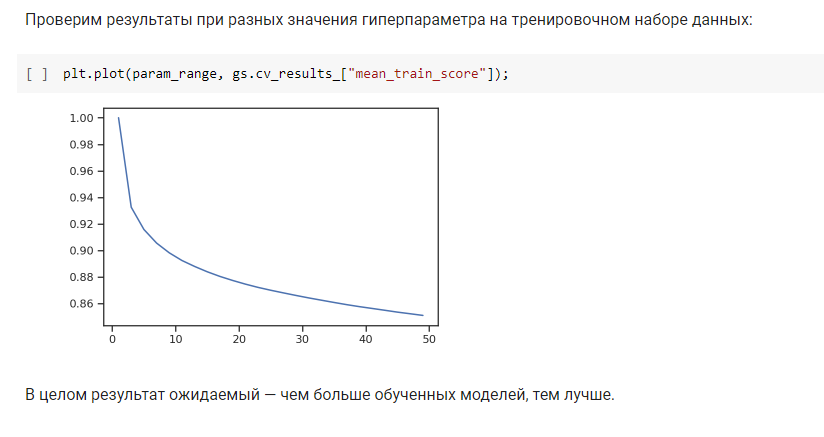
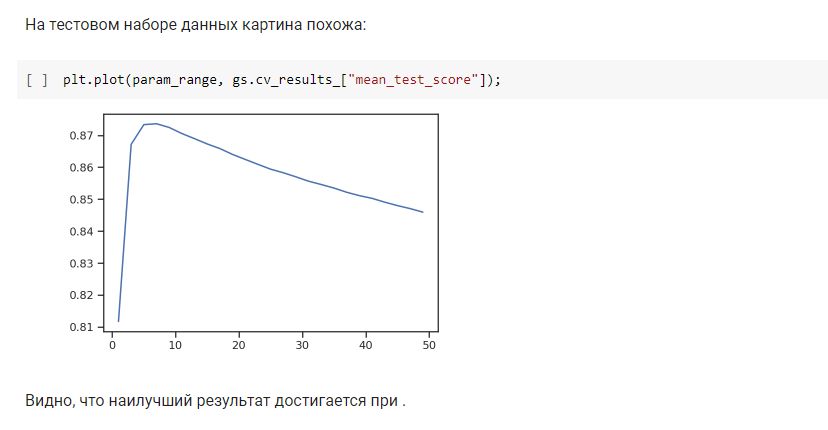
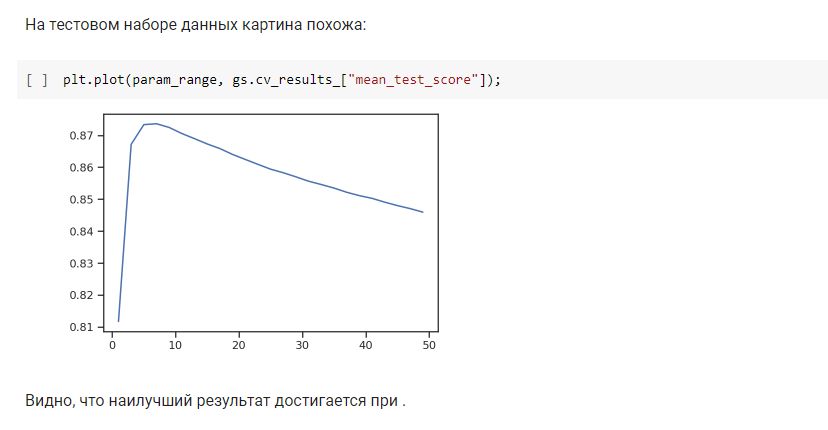


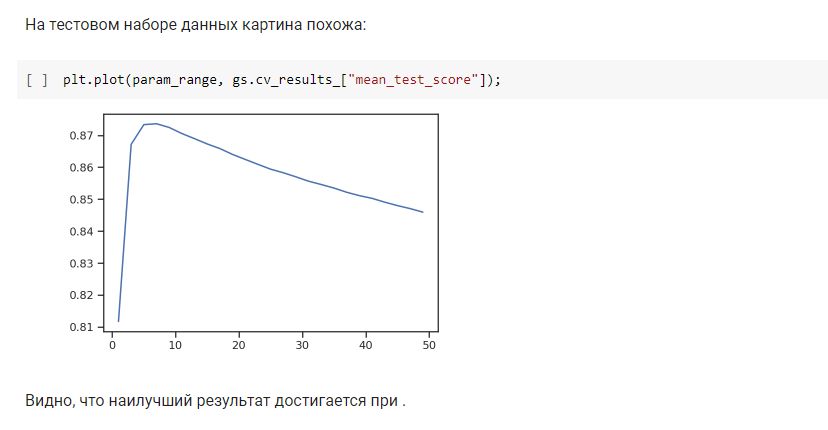
**2.9.1. Случайный лес**

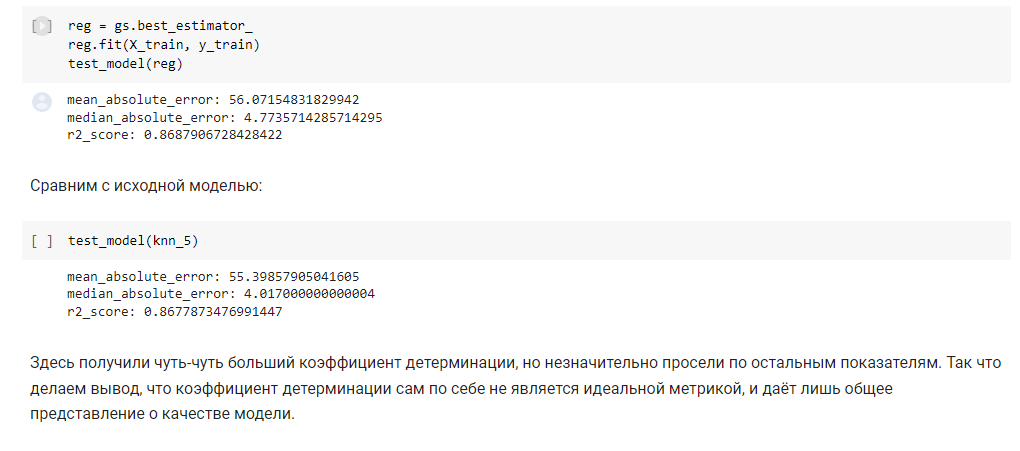


**2.10. Подбор гиперпараметров**  


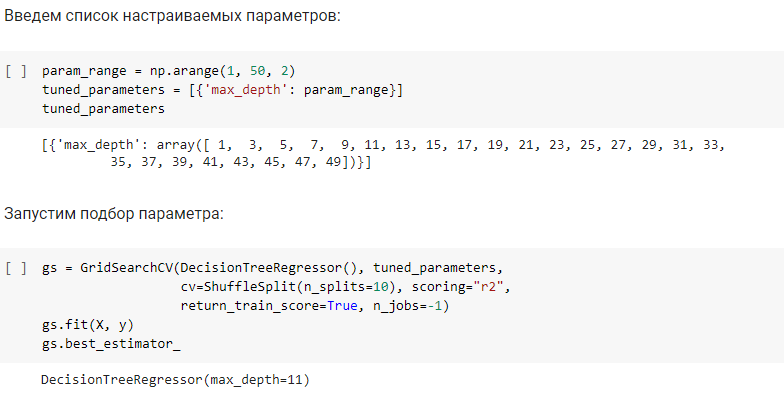
**2.10.1. Метод** *k* **ближайших соседей**  

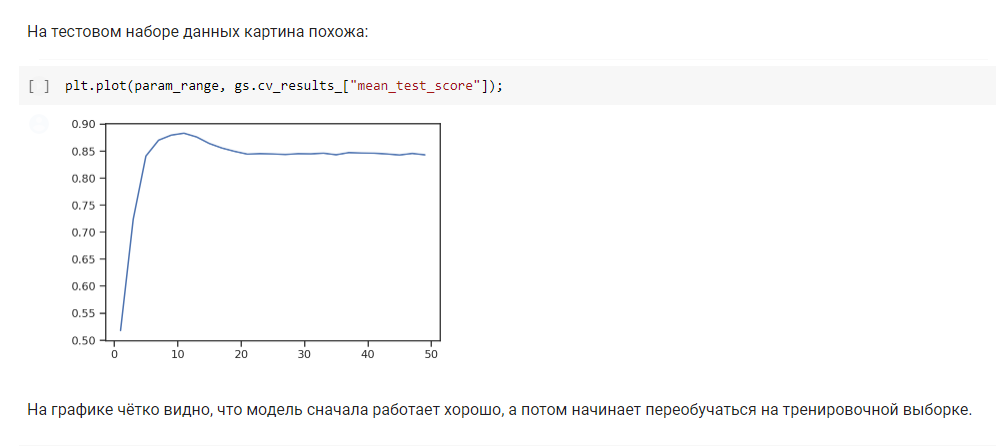


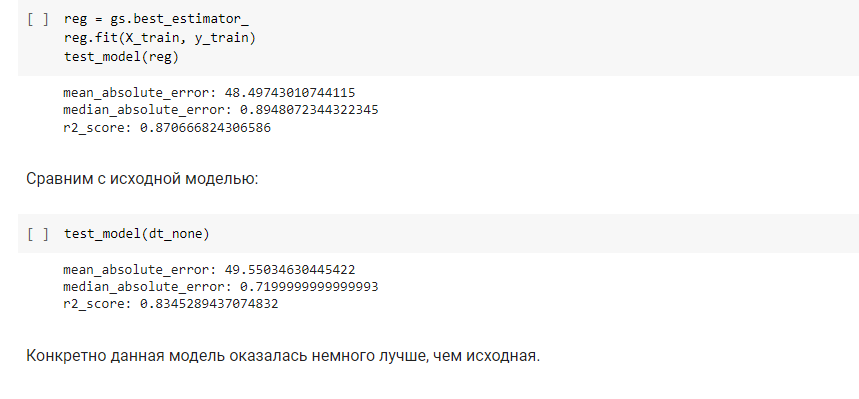


**2.10.2. Дерево решений**

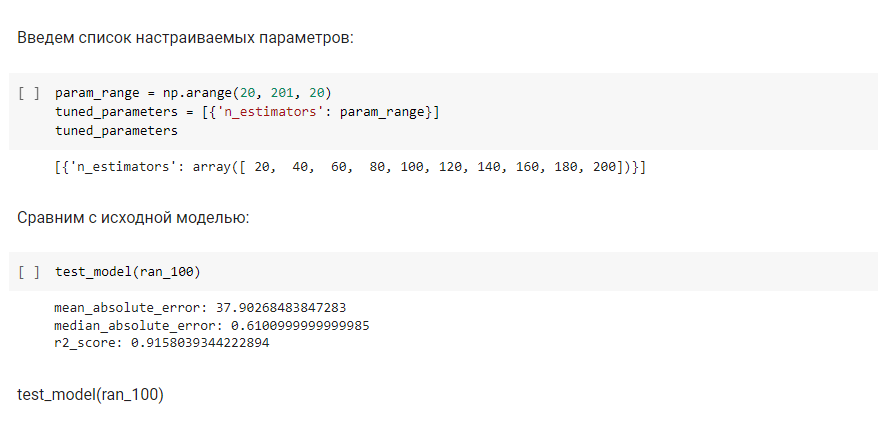








**2.10.3. Случайный лес**

  
**3. Выводы**  
 Все посмроенные модели обладают очень хорошими показателями. Ансамблевая модель при этом обладает наилучшими характеристиками. Таким образом для дальнейшей работы стоит использовать именно её.

**Список литературы**  
[1] Гапанюк Ю. Е. Домашнее задание по дисциплине «Методы машинного обучения»  
[Электронный ресурс] // GitHub. –– 2019. –– Режим доступа: https://github.com/  
ugapanyuk/ml\_course/wiki/MMO\_DZ (дата обращения: 06.05.2019).  
[2] You are my Sunshine [Electronic resource] // Space Apps Challenge. — 2017. —  
Access mode: https://2017.spaceappschallenge.org/challenges/earth-and-us/  
you-are-my-sunshine/details (online; accessed: 22.02.2019).  
[3] dronio. Solar Radiation Prediction [Electronic resource] // Kaggle. — 2017. — Access  
mode: https://www.kaggle.com/dronio/SolarEnergy (online; accessed: 18.02.2019).  
[4] Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource] //  
Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/  
stable/ (online; accessed: 20.02.2019).  
[5] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. —  
Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).  
[6] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode:  
http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).  
[7] Chrétien M. Convert datetime.time to seconds [Electronic resource] // Stack Overflow. — 2017. — Access mode: https://stackoverflow.com/a/44823381 (online; accessed: 20.02.2019).