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

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №2

«Методы построения моделей машинного обучения»

Вариант № 14

Выполнил: Проверил:  
Лупарев С.В. Гапанюк Ю.Е.  
группа ИУ5-63Б

Дата: 04.05.25 Дата:

Подпись: Подпись:

Москва, 2025 г.

**Задание:**

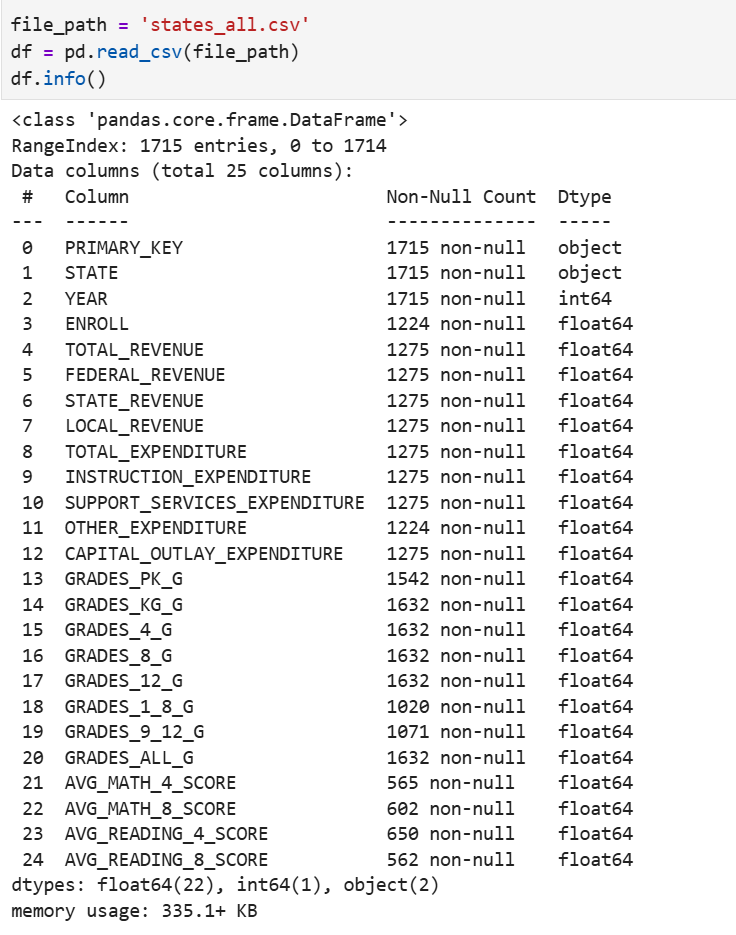
Номер варианта: **14**

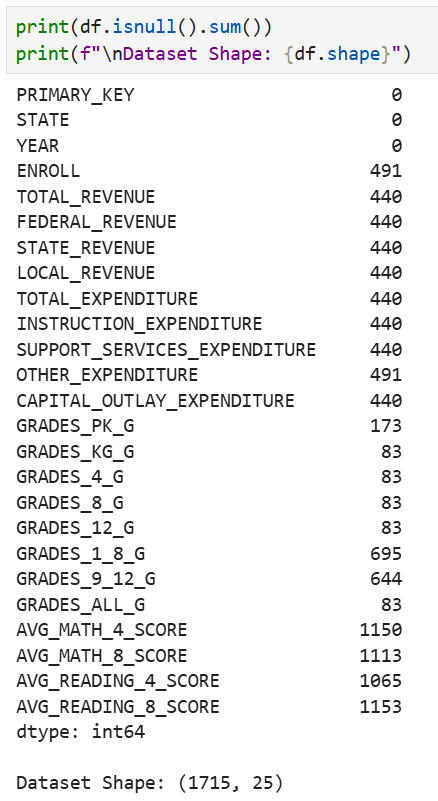
Номер набора данных, указанного в задаче: **14** https://www.kaggle.com/noriuk/us-education-datasets-unification-project (файл states\_all.csv)

Для заданного набора данных постройте регрессии. Для построения моделей используйте методы: Дерево решений, Случайный лес. Оцените качество моделей на основе подходящих метрик качества (не менее двух метрик). Какие метрики качества Вы использовали и почему? Какие выводы Вы можете сделать о качестве построенных моделей? Для построения моделей необходимо выполнить требуемую предобработку данных: заполнение пропусков, кодирование категориальных признаков, и т.д.

**Ход выполнения:**

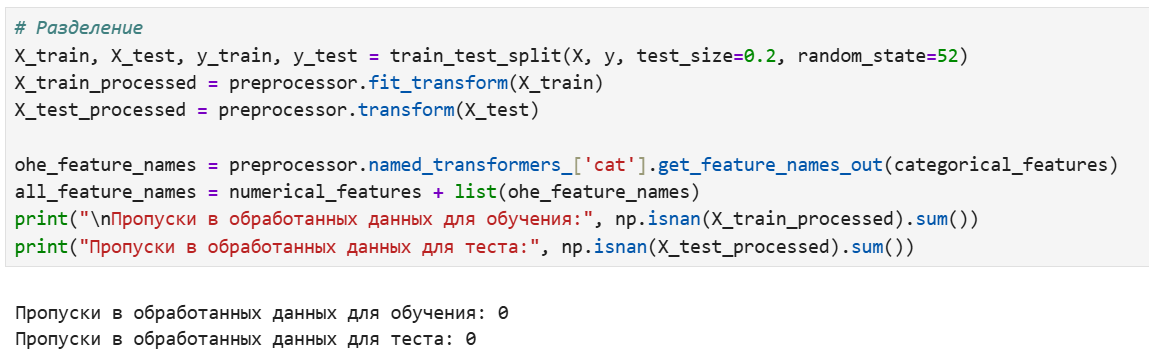
1. Загрузим данные, посмотрим количество пропусков





1. Будем предсказывать AVG\_MATH\_8\_SCORE, удалим все столбцы, связанные с оценками, удалим все пропуски, в которых AVG\_MATH\_8\_SCORE не определено, для остальных колонок пропуски заполним медианой для числовых и самым частым значением для категориальных. Также разделим данные на train и test.





1. Подберем лучшие параметры с помощью кросс-валидации для дерева и для случайного леса.

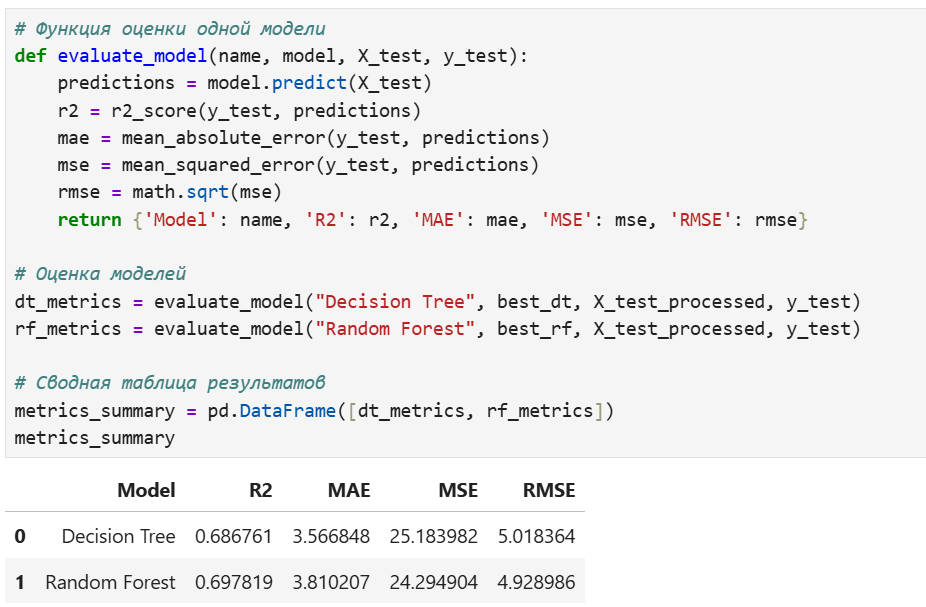




Результаты: Лучшие параметры для дерева: {'regressor\_\_max\_depth': 30, 'regressor\_\_max\_features': 1.0, 'regressor\_\_min\_samples\_leaf': 1, 'regressor\_\_min\_samples\_split': 20}

Лучшие параметры для леса: {'regressor\_\_max\_depth': 10, 'regressor\_\_max\_features': 0.7, 'regressor\_\_min\_samples\_leaf': 1, 'regressor\_\_min\_samples\_split': 2, 'regressor\_\_n\_estimators': 300}

1. Запустим модели с лучшими подобранными параметрами для наших test наборов данных и посчитаем метрики.



**Выводы:**

Для оценки качества моделей я использовал основные 4 метрики. Результаты у моделей оказались очень близкими, значение R2 нельзя назвать высоким, однако если мы посмотрим на MAE, то увидим, что оно немного меньше 4 для обеих моделей, при том, что значения предсказанной колонки в исходном наборе данных варьируются от 231 до 301, соответственно точность предсказаний моделей достаточно высокая.