**Министерство науки и высшего образования Российской Федерации** **Федеральное государственное бюджетное образовательное учреждение высшего образования** **«Московский государственный технический университет** **имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

**Факультет «Информатика и системы управления»**

**Кафедра ИУ5 «Системы обработки информации и управления»**

Отчёт по рубежному контролю №2

«Технологии машинного обучения»

Вариант 7

Выполнил:

студент группы ИУ5-63Б

Климов Н.С.

Преподаватель:

Гапанюк Ю. Е.

2024 г.

**Задание:** Для заданного набора данных (по Вашему варианту) постройте модели классификации или регрессии (в зависимости от конкретной задачи, рассматриваемой в наборе данных). Для построения моделей используйте методы 1 и 2 (по варианту для Вашей группы). 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.

Метод №1: Дерево решений

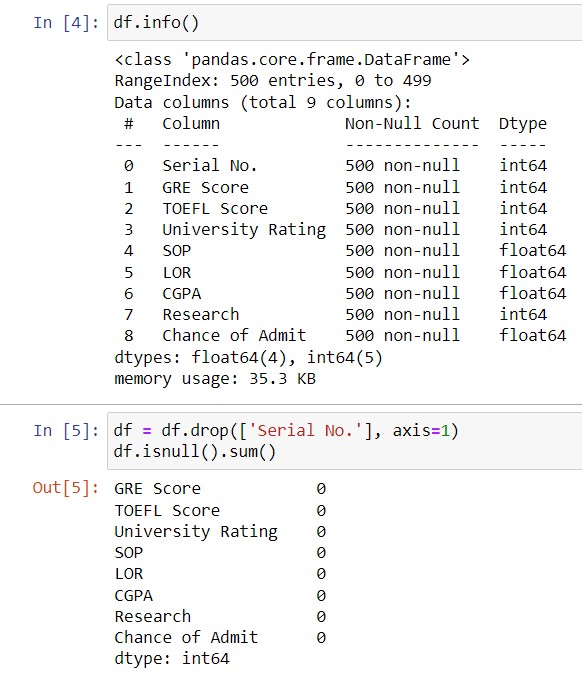
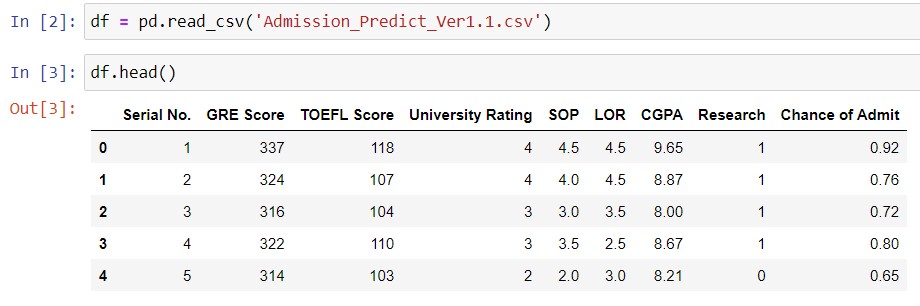
Метод №2: Случайный лес

Датасет: <https://www.kaggle.com/mohansacharya/graduate-admissions>(файл Admission\_Predict\_Ver1.1.csv)

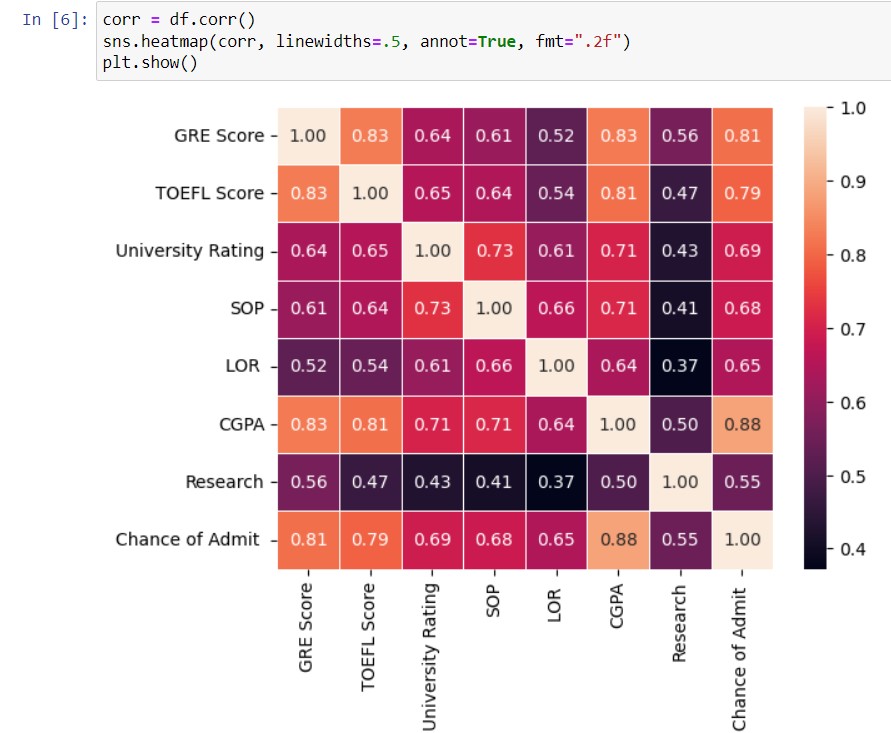
**Решение:**

Импортируем необходимые модули и загружаем данные из файла 'Admission\_Predict\_Ver1.1.csv' в объект `df`. Выводим первые строки данных с помощью `head()` и получаем информацию о данных с помощью `info()`. Удаляем столбец 'Serial No.' и подсчитываем количество пропущенных значений с помощью `isnull().sum()`.



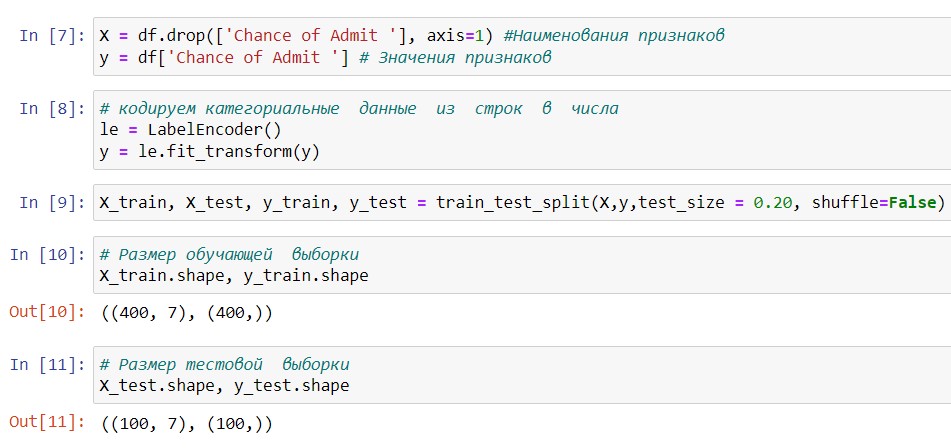


Создаём матрицу корреляции и визуализируем её с помощью тепловой карты, чтобы оценить связи между признаками в данных.

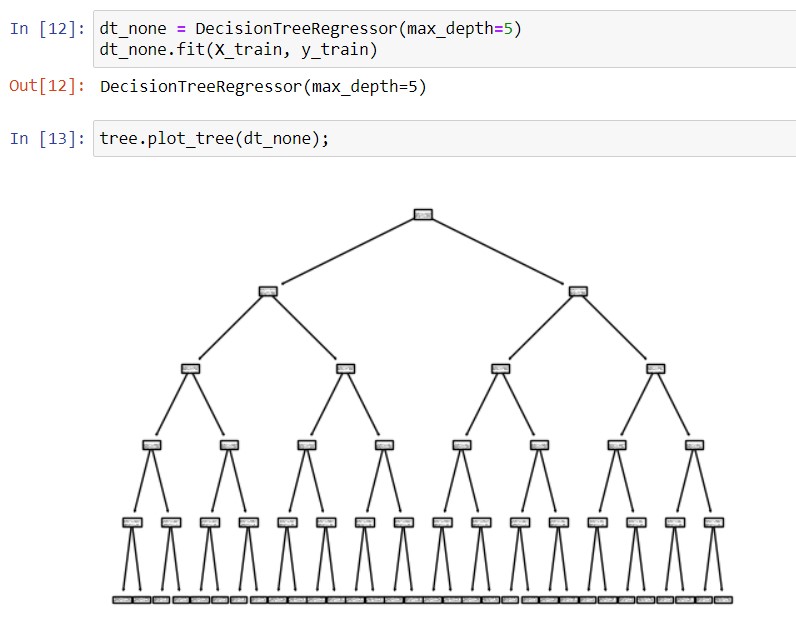


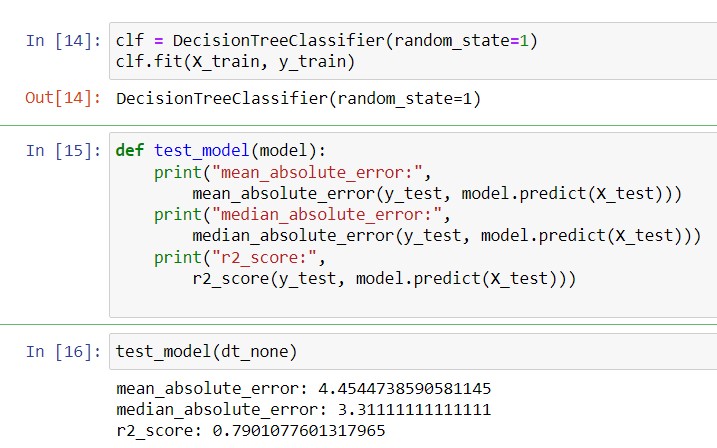
Наиболее сильная корреляция с целевым признаком "Chance of Admit" наблюдается у признаков "CGPA", "GRE Score" и "TOEFL Score". При построении модели машинного обучения эти признаки будут наиболее информативными и важными для прогнозирования результата. Следует отметить наличие корреляции между признаками "SOP" и "University Rating". Это может указывать на взаимосвязь между качеством заявки на поступление (SOP) и рейтингом университета, где студент подает заявку. Можно построить модель машинного обучения, используя признаки "CGPA", "GRE Score", "TOEFL Score", "LOR" и "Research". При этом особенно важными являются первые три признака, так как они сильно коррелируют с результатом.

Далее выполняем удаление столбца 'Chance of Admit', кодируем столбец 'Chance of Admit' из строковых значений в числовые и разбиваем данные на обучающую и тестовую выборки в соотношении 80/20.



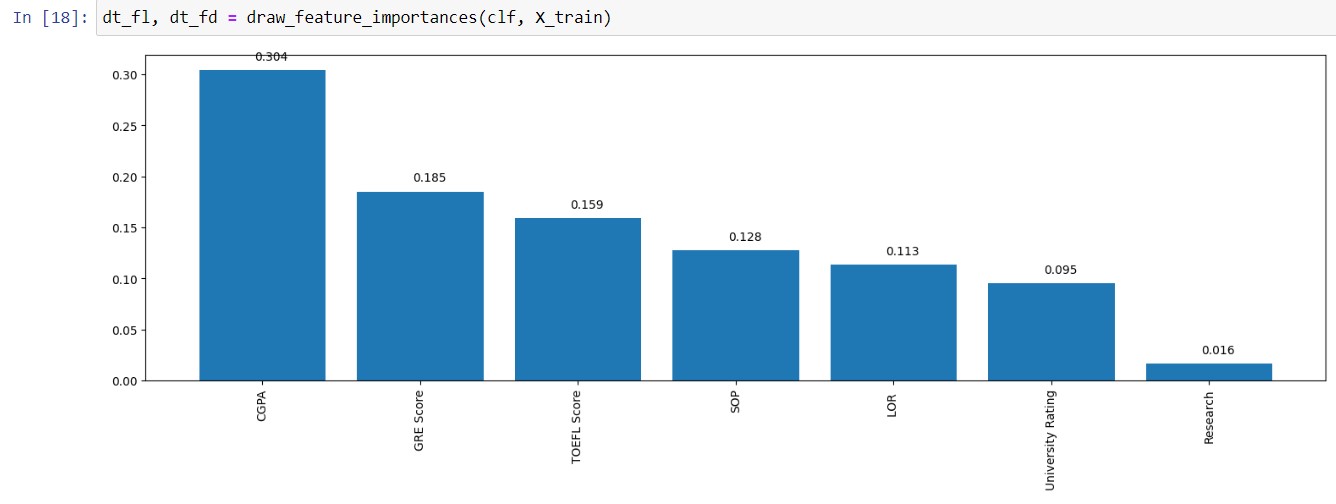
Затем создаем регрессор DecisionTreeRegressor с максимальной глубиной 5 и обучаем его на обучающих данных. Затем мы строим **дерево решений** для визуализации полученной модели. Далее, мы создаем классификатор DecisionTreeClassifier с указанным случайным состоянием и обучаем его на обучающих данных. Также определена функция test\_model() для оценки модели с использованием различных метрик. В конце вызываем функцию test\_model() для оценки модели dt\_none на тестовых данных.





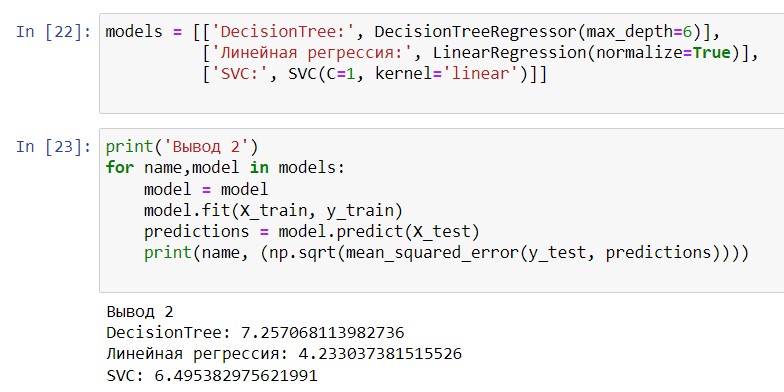
Определим функцию `draw\_feature\_importances`, которая выводит график важности признаков на основе модели дерева решений (`clf`) и набора данных признаков (`X\_train`). График отображает столбцы с названиями признаков и их важностью, а также выводит значения важности над соответствующими столбцами.





Далее выполняем поиск наилучших параметров для модели `DecisionTreeClassifier` с помощью кросс-валидации и оцениваем ее точность на тестовых данных, а также сравниваем ее с другими моделями.





После создаем модель **случайного леса** (`RandomForestClassifier`), обучаем ее на тренировочных данных, оцениваем ее точность на тестовых данных, а затем выполняем поиск наилучших параметров для модели с помощью кросс-валидации и оцениваем ее точность на тестовых данных с использованием найденных параметров.

