**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 4

Выполнила:

студентка группы ИУ5-63Б

Жамнова М.С.

Преподаватель:

Гапанюк Ю. Е.

2023 г.

Задание:

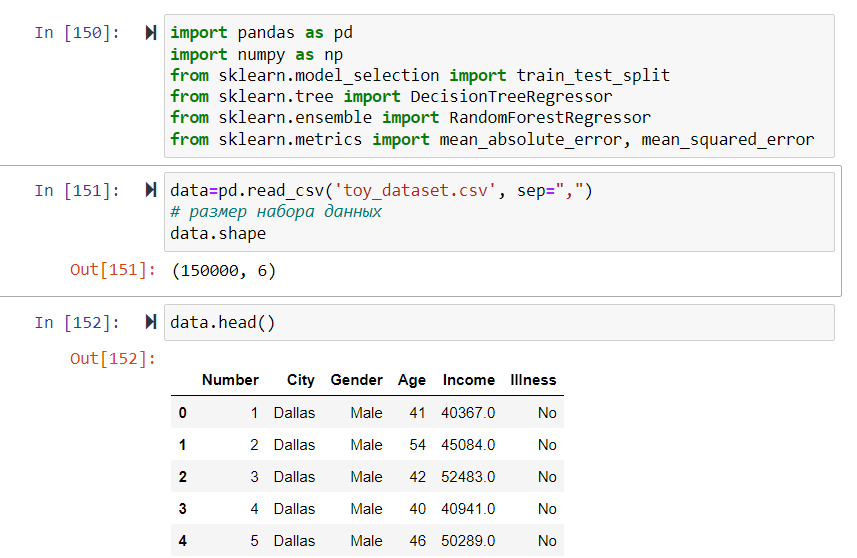
Для заданного набора данных (по Вашему варианту) постройте модели классификации или регрессии (в зависимости от конкретной задачи, рассматриваемой в наборе данных). Для построения моделей используйте методы 1 и 2 (по варианту для Вашей группы). Оцените качество моделей на основе подходящих метрик качества (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.

| **Группа** | **Метод №1** | **Метод №2** |
| --- | --- | --- |
| ИУ5-63Б, ИУ5Ц-83Б | Дерево решений | Случайный лес |

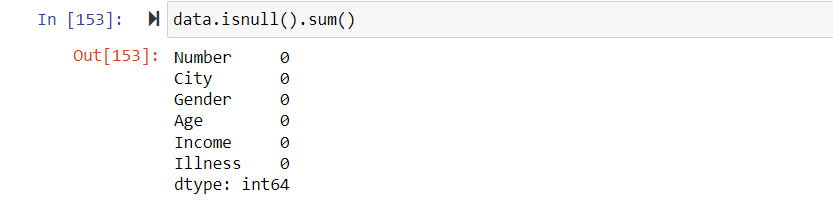
<https://www.kaggle.com/carlolepelaars/toy-dataset>

Решение:

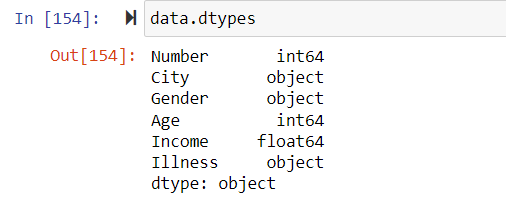
Загружаем датасет и подключаем необходимые библиотеки:



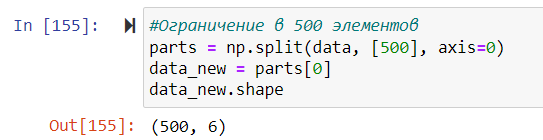
Посчитаем количество пустых значений:



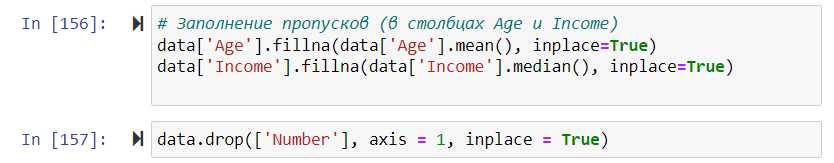
Проверка типов данных столбцов:



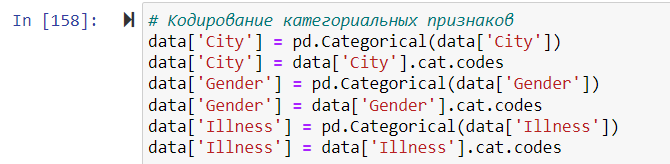
Ограничение данных до 500 строк:



Заполнение пропущенных значений и удаление не информативного столбца:

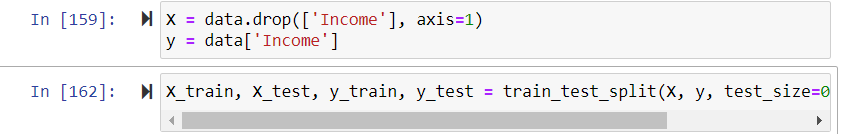


Кодирование категориальных признаков:



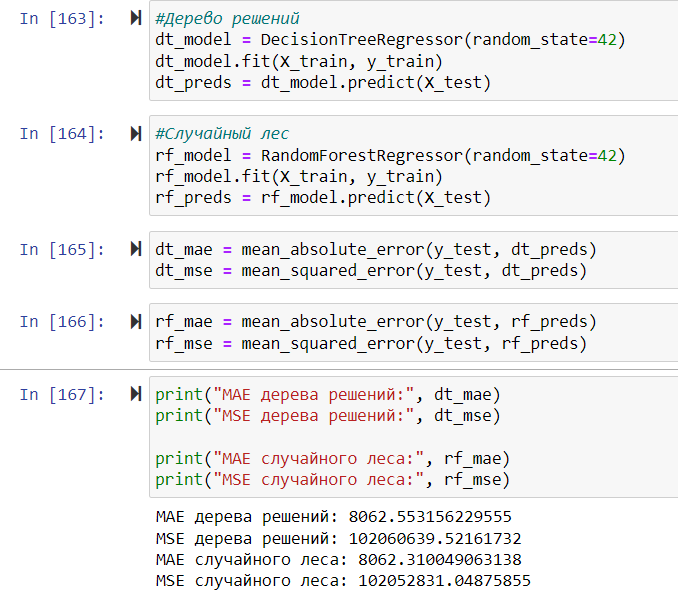
Мы преобразуем категориальные признаки в числовые значения, используя метод кодирования категорий. Каждое уникальное значение категориального признака заменяется числом, начиная с 0.

Разделение данных на обучающую и тестовую выборки:



Мы разделяем данные на обучающую и тестовую выборки в соотношении 70%/30% с использованием функции train\_test\_split. X представляет набор признаков, а y – целевую переменную.

Обучение моделей и оценка качества:



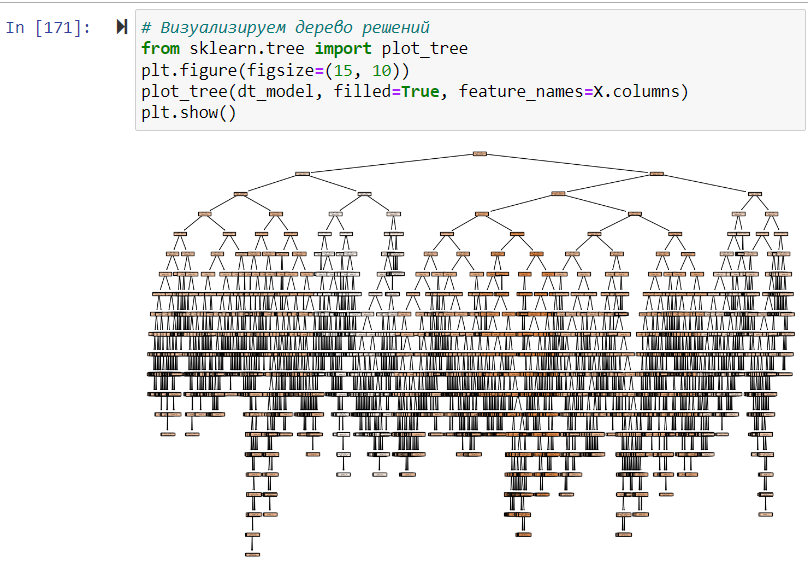
Cоздаются две модели - модель дерева решений и модель случайного леса. Каждая модель обучается на обучающих данных (X\_train и y\_train), а затем используется для получения прогнозов на тестовых данных (X\_test).

Для оценки качества прогнозов используются две метрики ошибки - средняя абсолютная ошибка (MAE) и средняя квадратичная ошибка (MSE). Для каждой модели рассчитываются значения этих метрик на тестовых данных.

Выводятся значения MAE и MSE для каждой модели. Интерпретация этих метрик зависит от контекста и задачи, которую решает модель. Обычно меньшие значения MAE и MSE свидетельствуют о лучшей точности прогнозов.

Визуализация дерева решений:

Мы визуализируем дерево решений с помощью функции plot\_tree из библиотеки scikit-learn. Мы используем параметр filled=True, чтобы закрасить узлы дерева в соответствии с прогнозируемой переменной, и передаем имена признаков как параметр feature\_names.



Вывод:

В данном примере были использованы две метрики качества - средняя абсолютная ошибка (MAE) и средняя квадратичная ошибка (MSE).

MAE - это средняя абсолютная разница между прогнозами и фактическими значениями. Она показывает, насколько сильно отклоняются предсказания модели от реальных значений. MAE является более понятной метрикой, так как ее значения можно интерпретировать как фактические расстояния между прогнозами и реальными значениями.

MSE - это средняя квадратичная ошибка между предсказаниями и фактическими значениями. Она показывает, насколько сильно отличаются прогнозы модели от реальных значений. MSE чувствительнее к большим ошибкам, так как она возводит разницу в квадрат.

Оценка качества моделей на тестовых данных показала, что метрики MAE и MSE для моделей дерева решений и случайного леса показывают результаты сравнимые по показателям качества. Это может свидетельствовать о том, что модели не сильно отличаются друг от друга по точности предсказаний. Однако для окончательных выводов необходима более обширная оценка, прежде всего, на больших объемах данных, чтобы подтвердить или опровергнуть эти выводы.