ЛАБОРАТОРНАЯ РАБОТА 6.

**Исследование набора данных**

**Цели и задачи**

Цель лабораторной работы: изучение принципов построения информационных систем с использованием линейных методов машинного обучения.

Основные задачи:

* освоение методологии работы с моделями линейной регресии в задачах машинного обучения;
* освоение метдик работы с линейными моделями в python;
* освоение методики применения методов регрессии;
* изучение основных параметров регрессионных моделей.

**Индивидуальное задание**

**Описание набора данных**

Данный набор данных является коллекцией пользовательских данных из социальных сетей. Он содержит информацию о отдельных пользователях, их паттернах использования и характеристиках.

Цель набора данных:

Целью этого набора данных является анализ поведения пользователей в социальных сетях, понимание их поведения, использования и выявление трендов. Этот набор данных может быть использован для построения моделей, предсказывающих вовлеченность пользователей, идентификации пользователей в сети и оптимизации маркетинга в социальных сетях.

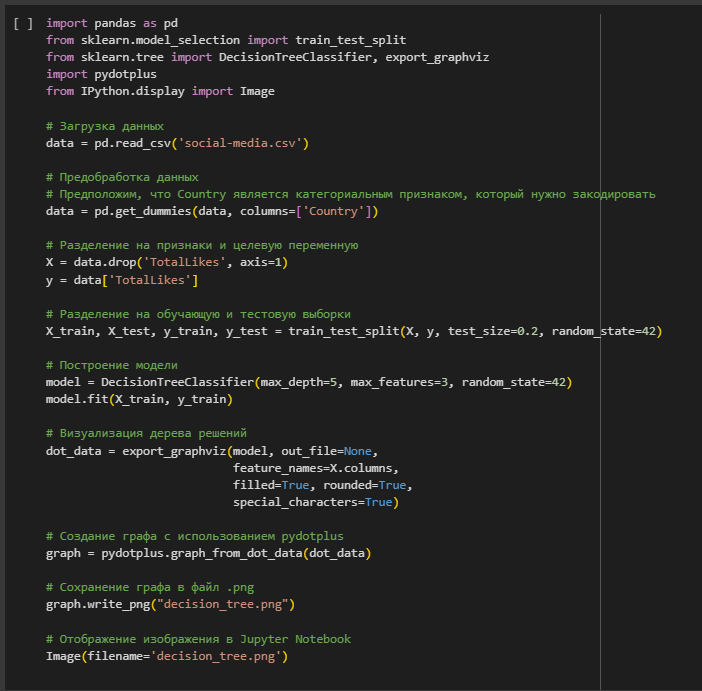
Описание каждого признака и его тип:

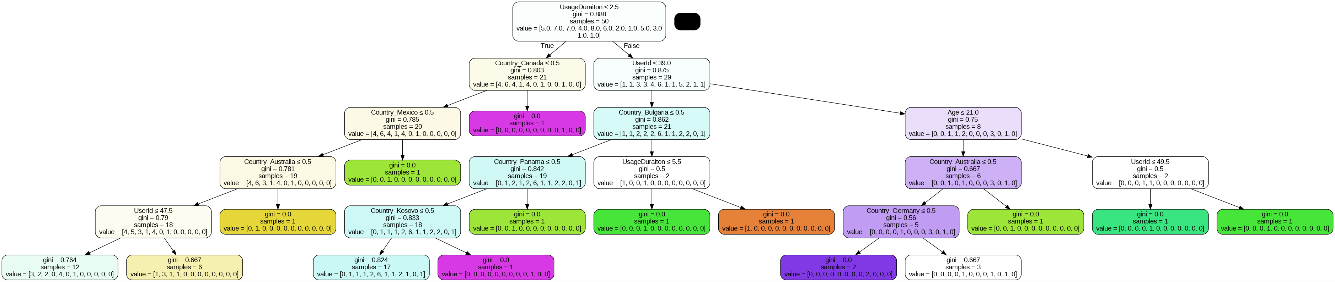
* UserId: Уникальный идентификатор для каждого пользователя в наборе данных (целое число)
* UsageDuration: Общее время, проведенное пользователем в социальных сетях, в часах (число)
* Age: Возраст пользователя в годах (число)
* Country: Страна проживания пользователя (категориальная переменная)
* TotalLikes: Общее количество лайков, поставленных пользователем за день (число)

**Индивидуальное задание**

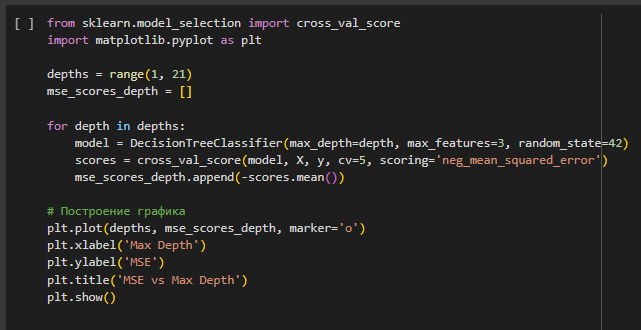
Выполните построение модели классификации на основе дерева классификации. В ходе решения задачи необходимо решить следующие подзадачи:

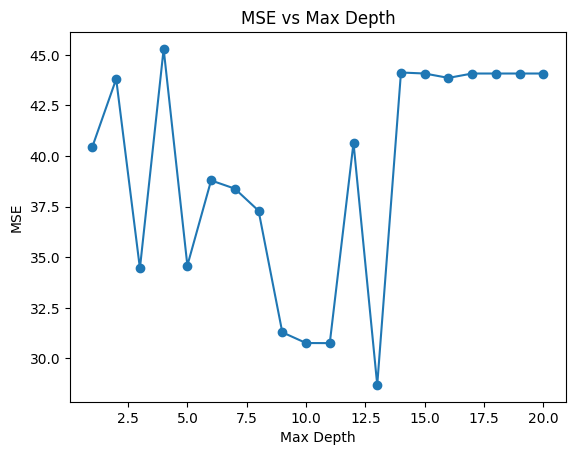
1 Построение логического классификатора с заданием max\_depth (максимальной глубины) и max\_features (максимального количества признаков) пользователем (установить любые); визуализация дерева решений для выбранных исследователем параметров (в формате .png)



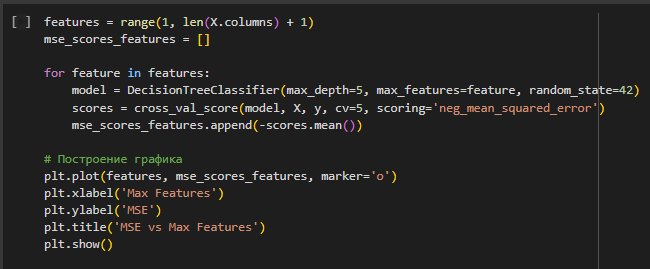


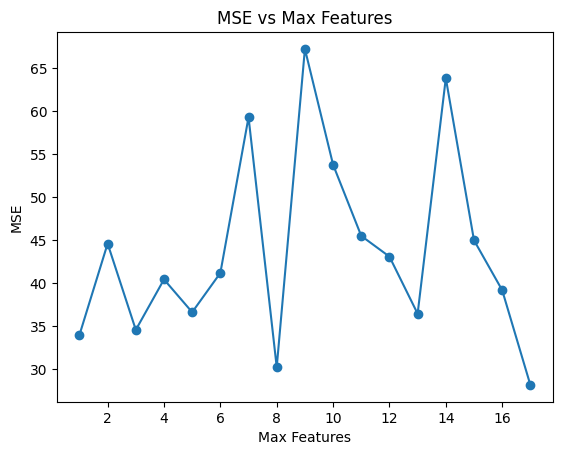
2 Вычисление оценки cross validation (MSE) для различнх значений max\_depth (построить график зависимости);



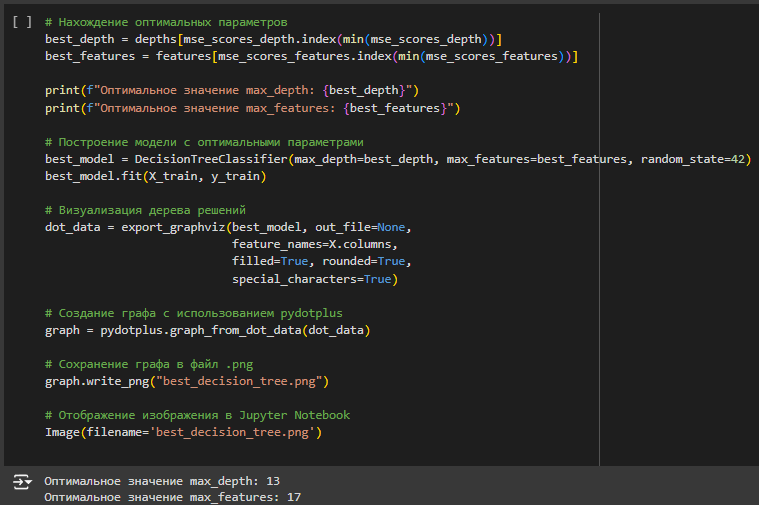


3 Вычисление оценки cross validation (MSE) для различнх значений max\_features (построить график зависимости);

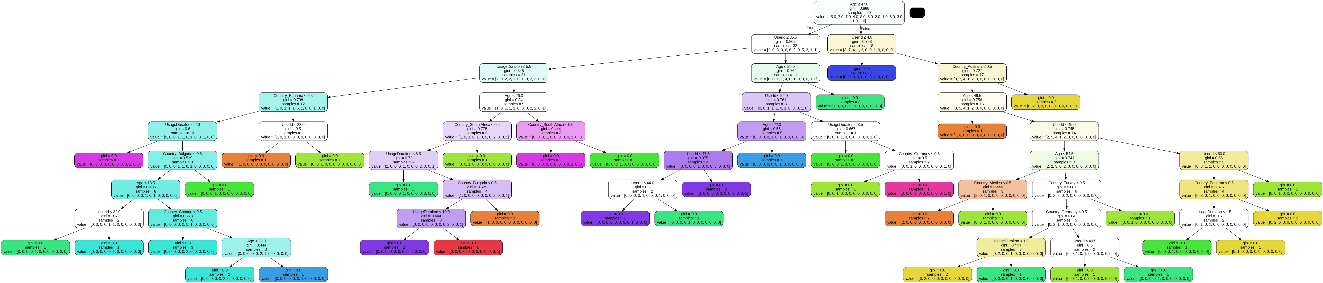




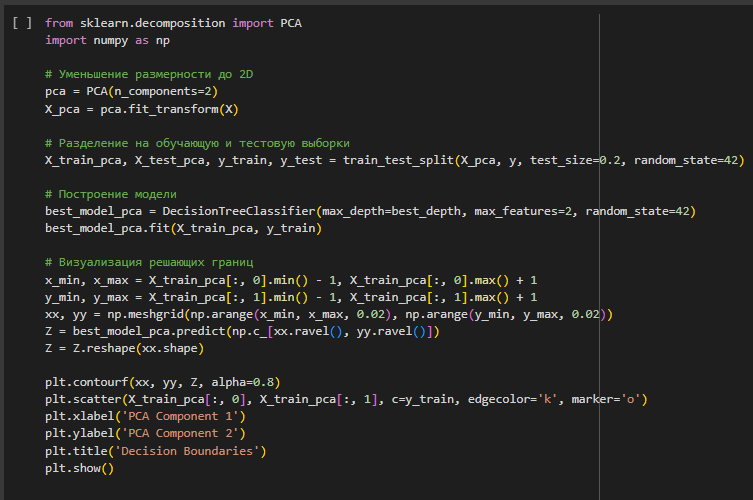
4 Вычислите оптимальные значения max\_depth и max\_features. Обоснуйте свой выбор. Продемонстрируйте использование полученного классификатора.

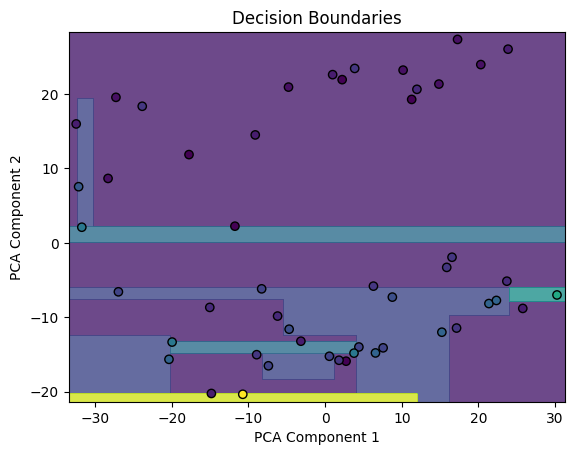


5 Выведите дерево в формате .png;



6 Выведите решающие границы полученной модели.





**Контрольные вопросы**

1. Какие методы классификации являются линейными?

Логистическая регрессия, Линейный дискриминантный анализ, Персептрон, Сupport Vector Machine с линейным ядром

1. Укажите основные параметры линеймой модели классификации.

Коэффициенты - Вектор коэффициентов, который определяет наклон гиперплоскости, Свободный член (bias или intercept), Регуляризация - Параметр, который контролирует сложность модели и предотвращает переобучение, Тип функции потерь: Например, кросс-энтропия для логистической регрессии или hinge loss для SVM.

1. Поясните назначение и принципы реализации методов стахостического градиента.

Назначение это - оптимизация функции потерь, используется для минимизации функции потерь в задачах обучения с учителем, и эффективность на больших наборах данных, он олее эффективен, чем обычный градиентный спуск, особенно на больших наборах данных.

Принципы реализации - итеративное обновление параметров: параметры модели обновляются на каждой итерации на основе градиента, вычисленного по одному случайно выбранному примеру (или небольшой подвыборке). Шаг обучения (learning rate): Определяет, насколько сильно параметры изменяются на каждой итерации. Сходимость: Процесс может быть остановлен, когда изменение функции потерь становится меньше определенного порога или достигнута максимальная количество итераций.

1. В чем заключается главная идея метода опорных векторов?

Главная идея метода опорных векторов (Support Vector Machine, SVM):

Максимизация зазора (margin): SVM стремится найти гиперплоскость, которая максимально разделяет классы, максимизируя зазор между классами.

Опорные векторы – Это точки, которые лежат на границе зазора или ближе всего к нему. Они играют ключевую роль в определении гиперплоскости.

Двойственная задача - SVM решает двойственную задачу оптимизации, которая позволяет использовать ядра для работы с нелинейно разделимыми данными.

1. Что такое «линейно разделимая выборка»?

Линейно разделимая выборка это когда выборка считается линейно разделимой, если существует гиперплоскость, которая может разделить все точки одного класса от всех точек другого класса без ошибок.

1. Поясните назначение ядер и спрямляющих пространств в алгоритмах линейной классификации.

Ядра используются для преобразования данных в более высокие пространства, где они могут стать линейно разделимыми.

Спрямляющие пространства — это пространства, в которые преобразуются исходные данные с помощью ядер.