ЛАБОРАТОРНАЯ РАБОТА 4. ЛОГИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

**Исследование набора данных**

**Цели и задачи**

Цель лабораторной работы: изучение принципов построения информационных систем с использованием логических методов классификации.

Основные задачи:

* освоение технологии внедрения алгоритмов на онове решающих списков в приложения;
* освоение технологии внедрения алгоритмов на онове решающих деревьев в приложения;
* изучение параметров логической классификации;
* освоение модификаций логических методов классификации.

**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.

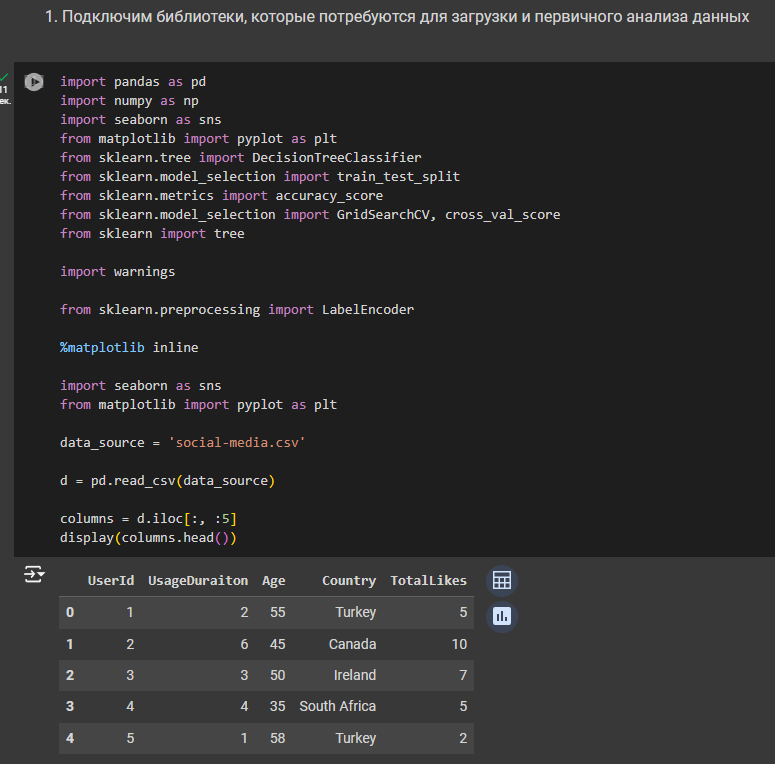
Цель набора данных:

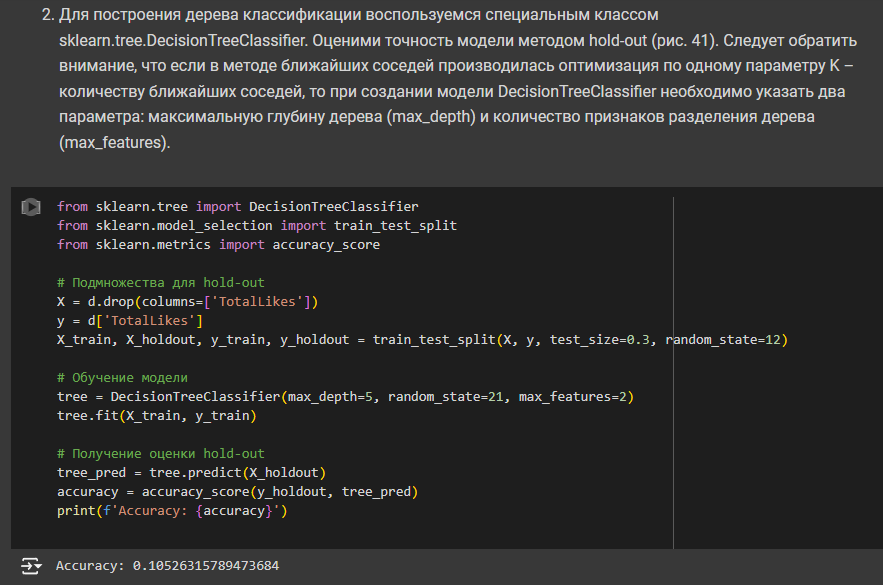
Целью этого набора данных является анализ поведения пользователей в социальных сетях, понимание их поведения, использования и выявление трендов. Этот набор данных может быть использован для построения моделей, предсказывающих вовлеченность пользователей, идентификации пользователей в сети и оптимизации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

Описание каждого признака и его тип:

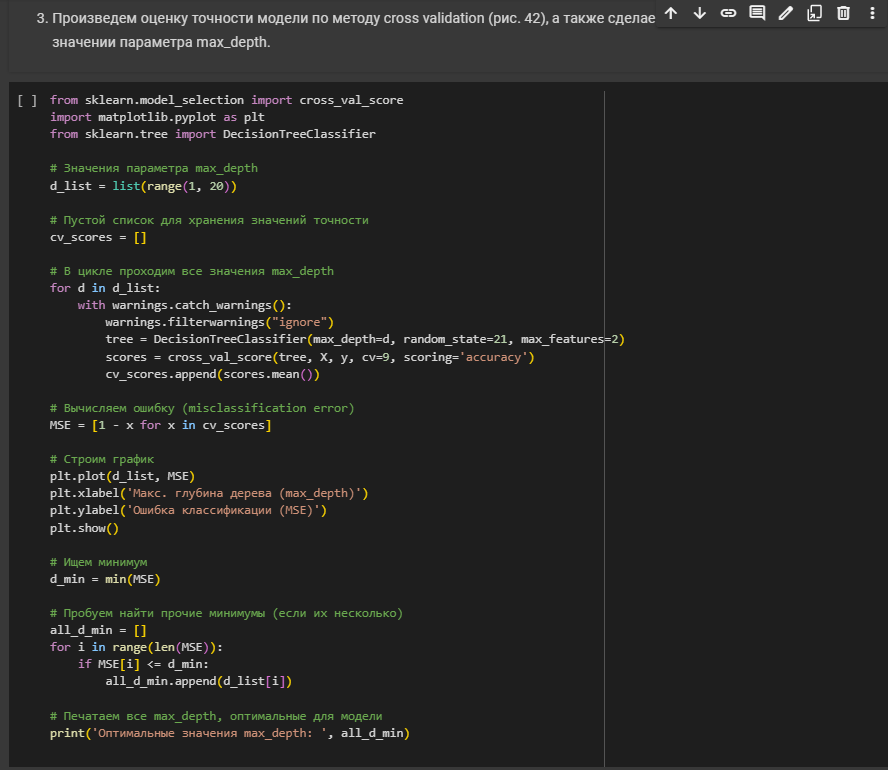
* UserId: Уникальный идентификатор для каждого пользователя в наборе данных (целое число)
* UsageDuration: Общее время, проведенное пользователем в социальных сетях, в часах (число)
* Age: Возраст пользователя в годах (число)
* Country: Страна проживания пользователя (категориальная переменная)
* TotalLikes: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

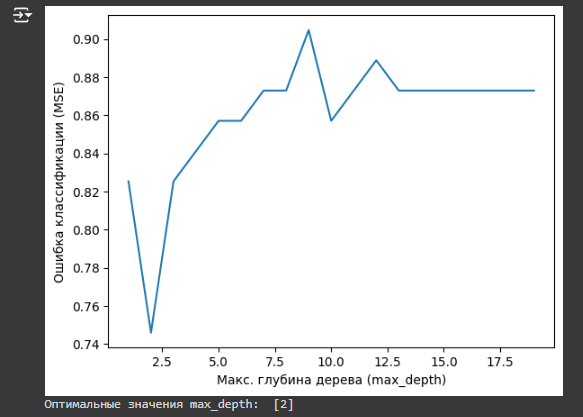
**Индивидуальное задание**

****

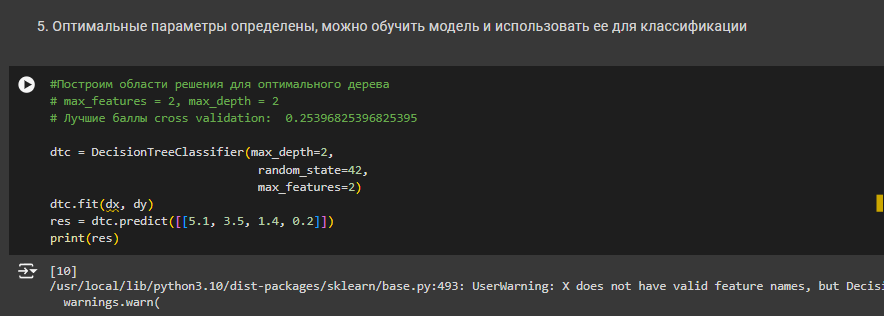
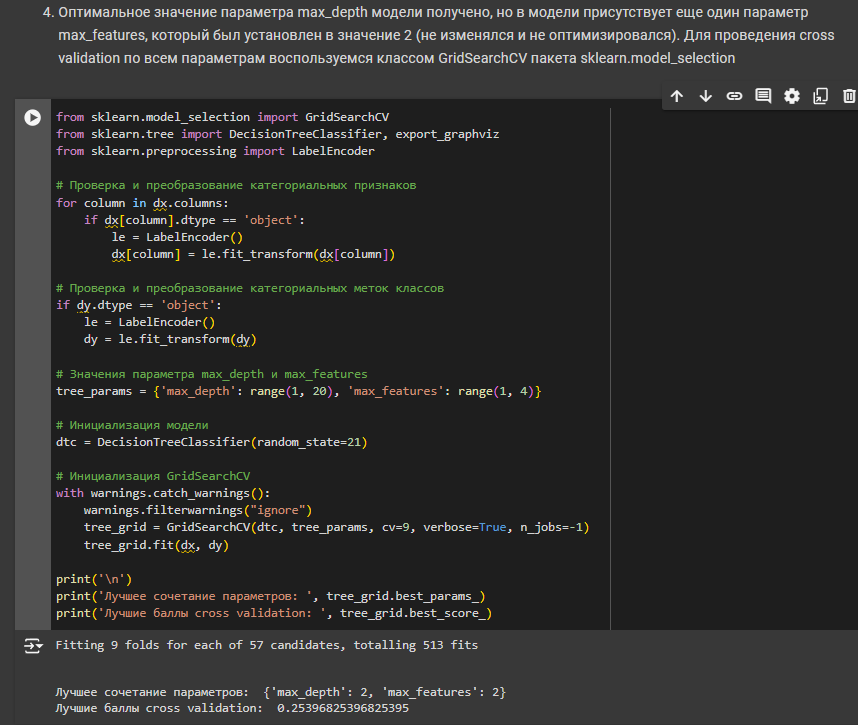
****

1,2) Построение логического классификатора с заданием max\_depth (максимальной глубины) и max\_features (максимального количества признаков) пользователем (установить любые); Вычисление оценки cross validation (MSE) для различнх значений max\_depth (построить график зависимости);

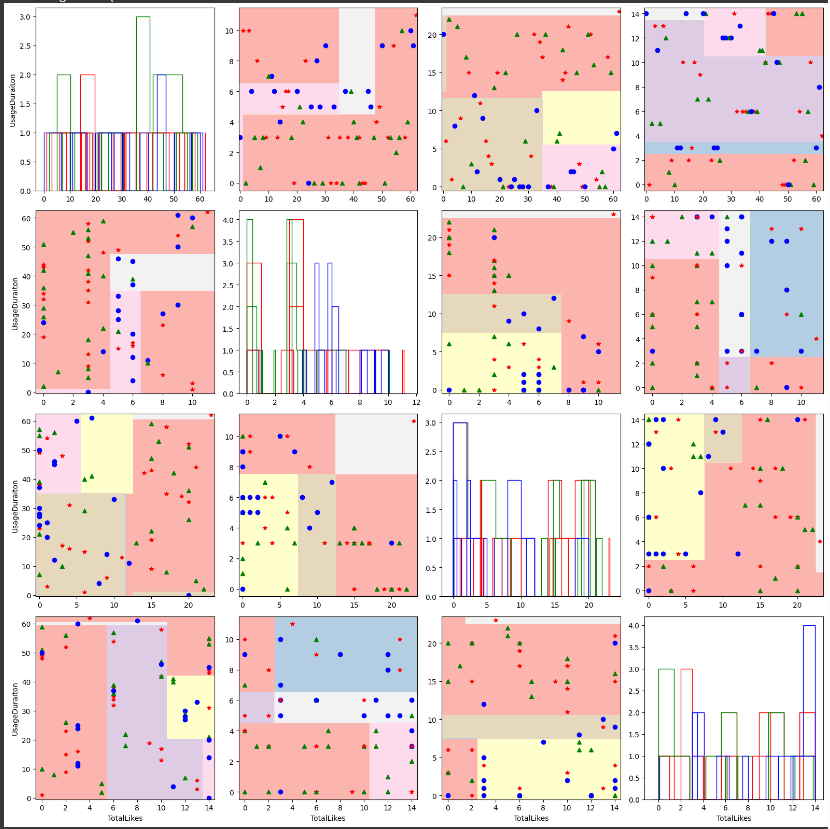
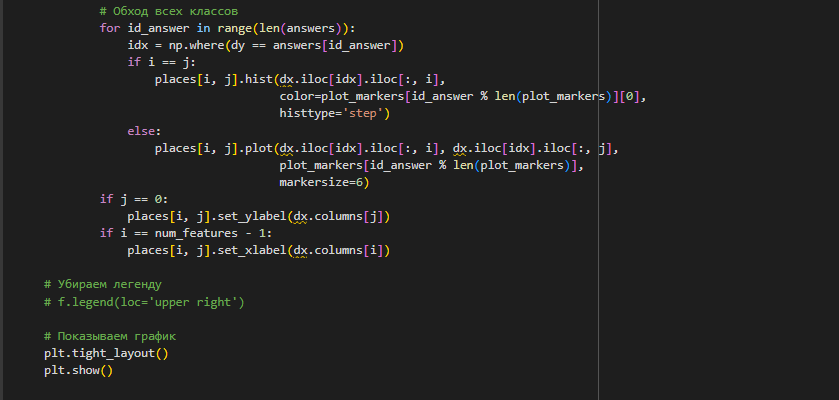
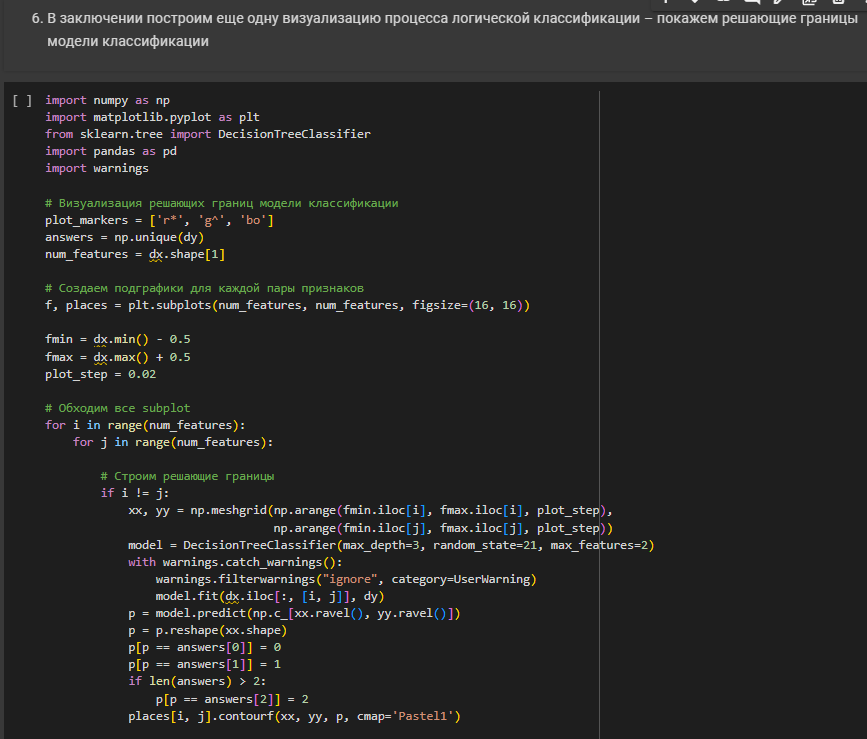




3, 4) Вычисление оценки cross validation (MSE) для различнх значений max\_features (построить график зависимости); Вычислите оптимальные значения max\_depth и max\_features. Обоснуйте свой выбор. Продемонстрируйте использование полученного классификатора.



1. Выведите решающие границы полученной модели.



**Контрольные вопросы**

1. Принцип построения дерева решений:

* Дерево решений строится рекурсивно, начиная с корня. На каждом шаге выбирается признак, который наилучшим образом разделяет данные на подмножества. Этот процесс продолжается до тех пор, пока все подмножества не станут однородными или не достигнут заданного критерия остановки

1. Статистическое определение информативности:

* Информативность признака определяется его способностью уменьшить неопределенность (или энтропию) в данных. Чем больше признак уменьшает энтропию, тем более информативным он считается.

1. Энтропийное определение информативности:

* Энтропия измеряет степень неопределенности в данных. Информативность признака определяется через информацию, которую он предоставляет, т.е. через уменьшение энтропии.

1. Многоклассовая информативность:

* Многоклассовая информативность используется для оценки признаков в задачах с несколькими классами. Она измеряет способность признака разделять данные на различные классы. Применяется для выбора наиболее важных признаков в многоклассовой классификации.

1. Назначение и алгоритм бинаризации количественных признаков:

* Назначение: Бинаризация используется для преобразования количественных признаков в бинарные, что может упростить модель и улучшить её интерпретируемость.

Алгоритм:

* Определяется пороговое значение.
* Все значения, превышающие порог, преобразуются в 1, а остальные в 0.

1. Порядок поиска закономерностей в форме конъюнкций:

Порядок:

* Определение целевой переменной и признаков.
* Выбор начальной конъюнкции (например, пустой).
* Рекурсивное добавление признаков в конъюнкцию, которые наилучшим образом улучшают качество модели (например, по критерию информативности).
* Остановка процесса, когда добавление новых признаков не улучшает качество модели или достигнут заданный критерий остановки.
* Оценка и выбор наилучшей конъюнкции.