Задание для четных вариантов. Случайный лес

1. Подготовить и проанализировать набор данных для обучения модели. Провести очистку данных и их предобработку (обработка пропусков, нормализация/стандартизация, кодирование категориальных признаков). Если необходимо выполнить масштабирование данных. Можно использовать датасет, который был выбран для разведывательного анализа данных.
2. Обучить модель случайного леса на выбранных данных: создать модель случайного леса с исходными параметрами. Обучить модель на тренировочной части данных. Сделать прогнозы на тестовой части.
3. Оценить качество модели с помощью метрик (accuracy, precision, recall, F1-score для классификации; MSE, RMSE, R^2 для регрессии).
4. Визуализировать результаты (например, матрицу ошибок для классификации).
5. Провести настройку гиперпараметров модели (число деревьев, глубина деревьев и т.п.).
6. Использовать GridSearchCV для подбора оптимальных параметров.
7. Повторить обучение и оценку с новыми параметрами.
8. Выполнить интерпретацию важности признаков.
9. Построить график важности признаков, определить, какие признаки вносят наибольший вклад в решение задачи.

Задание для нечетных вариантов. Градиентный бустинг

1. Подготовить и проанализировать набор данных для обучения модели. Провести очистку данных и их предобработку (обработка пропусков, нормализация/стандартизация, кодирование категориальных признаков). Если необходимо выполнить маштабирование данных. Можно использовать датасет, который был выбран для разведывательного анализа данных.
2. Обучить модель градиентного бустинга на выбранных данных: создать модель случайного леса с исходными параметрами. Обучить модель на тренировочной части данных. Сделать прогнозы на тестовой части.
3. Оценить качество модели с помощью метрик (accuracy, precision, recall, F1-score для классификации; MSE, RMSE, R^2 для регрессии).
4. Визуализировать результаты (например, матрицу ошибок для классификации).
5. Провести настройку гиперпараметров модели (число деревьев, глубина деревьев и т.п.).
6. Использовать GridSearchCV для подбора оптимальных параметров.
7. Повторить обучение и оценку с новыми параметрами.
8. Исследовать влияние скорости обучения (learning_rate) и количества деревьев на качество модели.