Лабораторная работа 3

Подбор гиперпараметров модели

Выполнял: Книга Тимофей
Группа: М80-309Б-23

Задание

- Выбрать модель для обучения (Desicion tree, Random forest, SVM, KNN, Boosting)

- Показать какие гиперпараметры есть у выбранной модели.

- Выбрать датасет для обучения и взависимости от модели подготовить данные

- Подобрать гиперпараметры для модели и сравнить лучшие подборы, для Grid Search, RandomSearch, Optuna

- На самом лучшем обучении сделать калькулятор, который показывает локальную интерпретацию с помощью LIME и глобальную интерпретацию с помощью SHAP

Данные

Датасет: Землетрясения (Kaggle)

Признак	Что означает	Диапазон / примечание
magnitude	Магнитуда землетрясения	Число (обычно 4–9)
cdi	Индекс ощутимости (Community Determined Intensity)	Целое число (0–10)
mmi	Интенсивность Меркалли (Modified Mercalli Intensity)	Целое число (0–10)
sig	Значимость события (significance)	Число (0–1000+)
nst	Количество станций, зафиксировавших событие	Целое число
dmin	Минимальное расстояние до станции (degrees)	Число (0–10+)
gap	Угловой разрыв между станциями (degrees)	Число (0–360)
depth	Глубина очага (км)	Число (0–700+)
latitude	Широта	Число (−90..90)
longitude	Долгота	Число (−180..180)
Year	Год	2022 (все записи из 2022)
Month	Месяц	1–12
tsunami	Был ли цунами (целевая переменная)	0 = нет, 1 = да

Что такое гиперпараметры

- Гиперпараметры — это настройки модели, которые мы выбираем ДО обучения (в отличие от “весов”, которые учатся сами).

- Пример для дерева решений:
    - глубина дерева (max_depth),
    - сколько объектов нужно для разделения (min_samples_split),
    - критерий качества (gini/entropy).

- Если гиперпараметры слишком “сильные”, модель переобучится: отлично запомнит обучение, но хуже будет работать на новых данных.

- Подбор гиперпараметров помогает найти баланс: хорошее качество + стабильная работа на тестовых данных.

Какие методы подбора мы использовали

- Grid Search:
  - перебираем ВСЕ комбинации параметров из сетки. Надёжно, но может быть долго.

- Random Search:
  - пробуем СЛУЧАЙНЫЕ комбинации. Часто быстрее и даёт сравнимый результат.

- Optuna (байесовская оптимизация):
  - “умный поиск”, который учится на прошлых попытках и предлагает более перспективные варианты.

- TPOT (AutoML):
  - автоматически подбирает не только гиперпараметры, но и саму модель + шаги предобработки.

Grid Search

Grid Search перебирает заранее заданную “сетку” значений и находит лучшую комбинацию.

Параметр	Лучшее значение
classifier__criterion	entropy
classifier__max_depth	5
classifier__max_features	None
classifier__min_samples_leaf	4
classifier__min_samples_split	2

Лучшее качество на кросс-валидации: 0.9136

Точность на тесте: 0.9427

Random Search

Random Search делает ограниченное число случайных попыток — обычно это быстрее, чем полный перебор.

Параметр	Лучшее значение
classifier__min_samples_split	5
classifier__min_samples_leaf	4
classifier__max_features	None
classifier__max_depth	7
classifier__criterion	gini

Лучшее качество на кросс-валидации: 0.9072

Точность на тесте: 0.9427

Optuna (байесовская оптимизация)

Optuna выбирает следующие варианты параметров “умнее”, опираясь на результаты предыдущих попыток.

Параметр	Лучшее значение
classifier__max_depth	17
classifier__min_samples_split	3
classifier__min_samples_leaf	6
classifier__criterion	gini

Лучшее качество на кросс-валидации: 0.8928

Точность на тесте: 0.8981

Примечание: в этом запуске Optuna дала качество хуже, чем Grid/Random — возможно, нужно больше trial-ов или другая настройка диапазонов.

TPOT (AutoML)

- TPOT автоматически попробовал разные модели и шаги предобработки и выбрал лучший вариант по качеству.

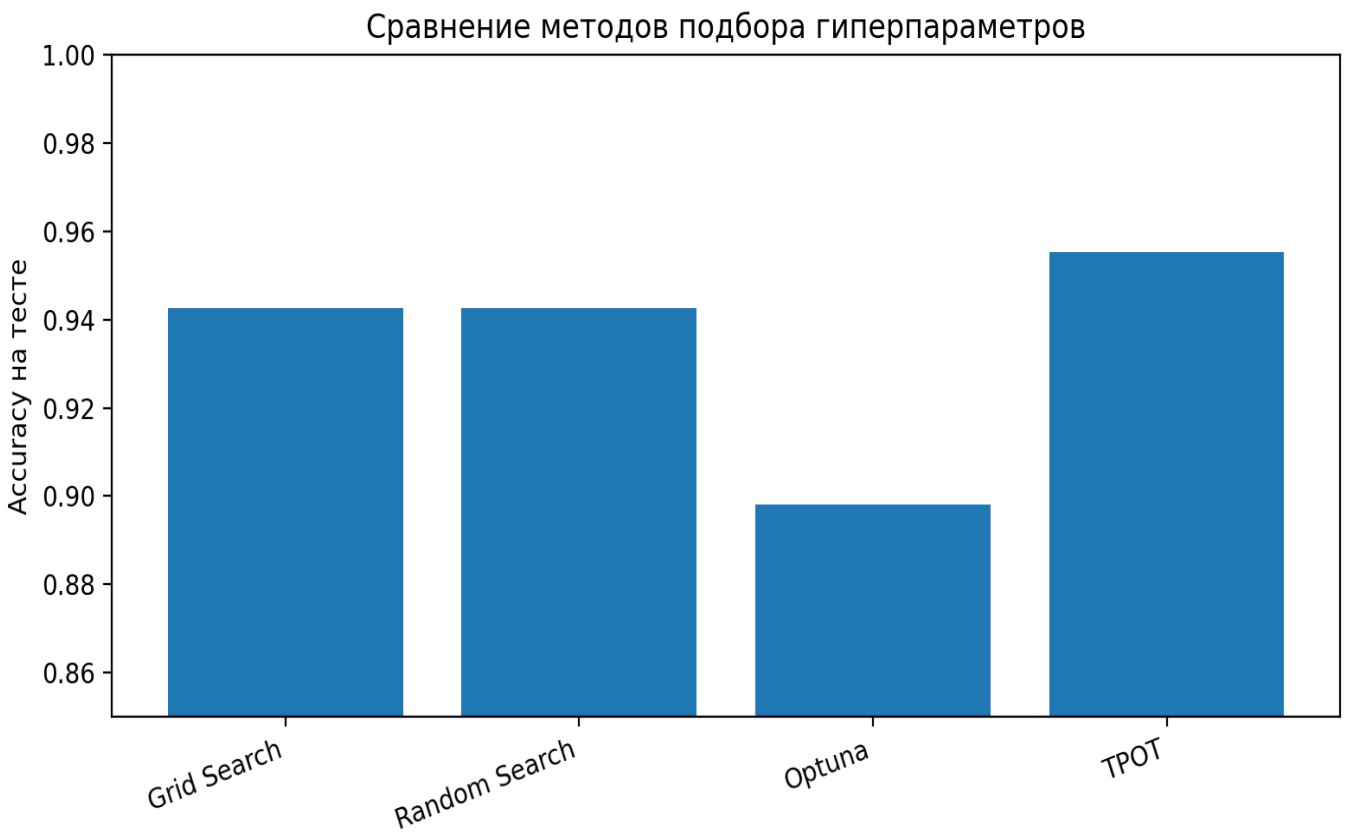
- Итоговый конвейер (упрощённо): MinMaxScaler → VarianceThreshold(threshold≈0.0160) → LightGBMClassifier.

- Параметры модели (из вывода): max_depth≈8, n_estimators≈90, num_leaves≈95.

- Точность на тесте: 0.9554 (лучшая среди всех методов в лабораторной работе).

Важно: TPOT сравнивает уже другие модели, поэтому результат не “чисто” про гиперпараметры одного дерева, а про поиск лучшего решения целиком.

Сравнение результатов



Метод	Accuracy (CV)	Accuracy (test)
Grid Search	0.9136	0.9427
Random Search	0.9072	0.9427
Optuna	0.8928	0.8981
TPOT	—	0.9554

Если сравнивать именно дерево решений, Grid и Random дали одинаковую точность на тесте.
TPOT дал лучшее качество, но за счёт выбора другой модели.

Выводы

- Подбор гиперпараметров действительно влияет на качество: “удачные” настройки дают заметный прирост по accuracy.

- Для DecisionTreeClassifier в этой работе Grid Search и Random Search показали лучший и одинаковый результат на тесте (≈0.9427).

- Optuna в текущем запуске сработала хуже — вероятно, из-за ограниченного числа попыток и/или неудачных диапазонов.

- TPOT (AutoML) показал максимум (≈0.9554), потому что смог выбрать другой алгоритм (градиентный бустинг деревьев) и предобработку.