Домашнее задание #1

Основы ML: EDA, модели, метрики, гипотезы

Курс ML Start 2026

Дедлайн: пятница, 13.02.2026, 23:59

Суть задания

Всё просто: берёте датасет, ставите гипотезы, проверяете их экспериментами. Никаких шаблонов – делайте как считаете нужным, главное чтобы было видно ход мысли.

Что нужно сделать

1. Выбрать один датасет из списка ниже (или предложить свой, но согласуйте).
2. Сформулировать 5–7 гипотез до начала работы. Гипотеза – это конкретное утверждение, которое можно проверить экспериментом. Примеры:
   - «Логистическая регрессия будет работать лучше дерева решений, потому что признаки линейно разделимы»
   - «Удаление коррелированных признаков (корреляция > 0.9) улучшит R^2 линейной регрессии»
   - «Ridge с подобранным λ обойдёт OLS на тесте из-за мультиколлинеарности»
   - «Признаки X, Y, Z будут самыми важными по feature importance дерева»
   - «При глубине дерева > 10 начнётся переобучение (train accuracy вырастет, test – упадёт)»
   - «F1 macro будет сильно отличаться от accuracy из-за дисбаланса классов»
   - «Отбор 5 лучших признаков через SelectKBest не ухудшит качество модели»
3. Провести работу – EDA, подготовка данных, обучение моделей, подбор гиперпараметров, сравнение.
4. По каждой гипотезе написать вывод: подтвердилась / не подтвердилась / частично, и почему.

Датасеты на выбор

Ниже – ссылки на датасеты. Все достаточно сложные и интересные, с разными типами задач.

Если хотите взять свой датасет – ок, но он должен быть не из числа «заезженных» (Titanic, Iris, Boston Housing, MNIST и т. д.) и содержать хотя бы 1000 строк.

Датасет	Задача	Размер	Ссылка
Steel Plates Faults	Классификация повреждений (7 типов дефектов)	1.9k	https://www.kaggle.com/datasets/eouedraogo4/steel-plates-faults
Obesity Levels	Классификация (7 классов)	2.1k	https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels
Rain in Australia	Бинарная классификация	145k	https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package
Mushroom Classification	Бинарная классификация	8.1k	https://www.kaggle.com/datasets/uciml/mushroom-classification
Online Shoppers Intention	Бинарная классификация	12.3k	https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset
Used Car Prices	Регрессия	188k	https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset

Что должно быть в работе

Необязательно делать всё подряд, но вот чеклист того, что мы хотим увидеть:

- ✓ EDA: распределения, корреляции, пропуски, дисбаланс классов
- ✓ Подготовка данных: обработка пропусков, кодирование категорий, масштабирование
- ✓ Минимум 2–3 модели (например: LogReg + Decision Tree + Ridge/Lasso)
- ✓ Метрики: подходящие для задачи (не только accuracy)
- ✓ Кросс-валидация
- ✓ Подбор гиперпараметров (GridSearch или RandomizedSearch)
- ✓ Отбор признаков (хотя бы один метод)
- ✓ Гипотезы и выводы по каждой

Не надо писать простыни текста. Код + краткие выводы в markdown-ячейках – достаточно.

Формат сдачи

GitHub-репозиторий со следующей структурой (примерно):

✓ README.md

✓ research.ipynb

✓ скачанный датасет

✓ доп файлы если потребуется

README.md должен содержать:

- Какой датасет выбрали и почему
- Список гипотез (до начала работы)
- Краткие результаты (что подтвердилось, что нет)
- Ссылка на данные (если не лежат в репо)

Ссылку на репозиторий скинуть до дедлайна.

Оценивание

Оценивается не «правильность» гипотез (они могут не подтвердиться – это нормально), а:

- Осмысленность гипотез (не «модель будет работать», а конкретное утверждение)
- Качество экспериментов (корректный пайплайн, нет утечки данных)
- Выводы (не просто «$R^2 = 0.85$», а почему так и что это значит)
- Аккуратность кода и репозитория

Дедлайн: пятница, 13 февраля 2026, 23:59

По всем вопросам пишите – @artemovma