Чекліст для проведення дослідницького аналізу даних

Автор: Ганна Пилєва

Чекліст для проведення дослідницького аналізу даних (EDA) в контексті машинного навчання дозволяє систематизувати процес підготовки та аналізу даних перед побудовою прогностичних моделей. Ось перелік кроків, аби ви були певні, що нічого не пропустили:

1. Розуміння бізнес-задачі

- ☐ Визначити цілі проєкту та ключові питання, на які потрібно відповісти.
- ☐ Зрозуміти, як результати аналізу будуть використані.

2. Збір та інтеграція даних

- ☐ Зібрати необхідні набори даних.
- ☐ Перевірити повноту та якість джерел даних.
- ☐ Інтегрувати (з’єднати) дані з різних джерел, якщо необхідно.

3. Первинний огляд даних

- ☐ Оцінити розмір датасету та структуру даних (таблиці, часові ряди, зображення тощо).
- ☐ Визначити типи змінних в даних (числові, категоріальні тощо):
  - ☐ **Категоріальні змінні:** Визначення унікальних категорій і їх розподілу. Перевірка на наявність порядкових змінних, які можуть вимагати спеціального кодування.
  - ☐ **Числові змінні:** Аналіз розподілу (нормальність, асиметрія тощо) і потенційні кореляції з цільовою змінною. Виявлення ознак з високою варіативністю, які можуть бути корисними для моделі.

- [ ] **Часові змінні:** Розуміння часових рамок даних та їх можливий вплив на цільову змінну.

- [ ] Провести первинний огляд наявних ознак і їх потенційну доцільність для задачі.
  - [ ] Ідентифікація ознак з високим потенціалом впливу на цільову змінну.
  - [ ] Визначення ознак, які можуть бути менш важливими або надмірними, засновуючись на їх розподілі та взаємозв'язках.

4. Очищення даних

- [ ] Виявити та обробити пропущені значення.
- [ ] Виправити або видалити аномалії та викиди.
- [ ] Видалити дублікати.

5. Аналіз та візуалізація даних

- [ ] Провести аналіз розподілу кожної змінної.
- [ ] Візуалізувати розподіли та взаємозв'язки між ознаками за допомогою графіків і діаграм.
- [ ] Виявити кореляції між ознаками.

6. Підготовка ознак (Feature Engineering)

- [ ] Створення нових ознак на основі наявних даних.
- [ ] Вибір або відкидання ознак для моделювання.
- [ ] Кодування категоріальних змінних (наприклад, One-hot encoding).
- [ ] Нормалізація або стандартизація числових змінних.

7. Розділення даних

- [ ] Розділити дані на навчальні, валідаційні та тестові вибірки.

8. Побудова базових моделей (Baseline)

- [ ] Розробити прості моделі для встановлення базового рівня ефективності.

- [ ] Оцінити продуктивність базових моделей.

9. Ітеративне удосконалення

- [ ] Провести ітерації удосконалення моделей на основі отриманих результатів.