Из исходного датасета **убрали все “готовые” оценки** и оставили только **сырые признаки**:

- В **CSV остаются**:
    - `market_size_million` – примерный размер целевого рынка в миллионах пациентов (задаётся/берётся как внешняя эпидемиологическая оценка).
    - `competition_level` – примерное количество конкурирующих препаратов / аналогов на рынке.
- Из **CSV убраны**, но теперь **вычисляются алгоритмом**:
    - `expected_profit_score`
    - `success_probability`
    - `traditional_time_years`
    - `ai_time_years`

Как теперь считаются эти колонки

- **`expected_profit_score`**
  Считается внутри пайплайна как интегральная оценка «потенциальной прибыльности», например на основе:
    - `market_size_million` (больше рынок → лучше),
    - `competition_level` (меньше конкурентов → лучше),
    - `efficiency_score` (эффективность),
    - `toxicity_score` (безопасность),
    - `uniqueness_score` (насколько кандидат отличается от существующих).
- **`success_probability`**
  Тоже не хранится в датасете, а считается по ходу алгоритма как оценка **вероятности успешного прохождения фаз I–III**.
  Используются те же признаки: эффективность, токсичность, уникальность,

размер рынка и конкуренция.

- **`traditional_time_years`**
  Это не “руками заданное число”, а **предсказание отдельной модели**:
  - обучаем отдельный регрессионный алгоритм на **своём датасете**, где для многих исторических примеров известны:
    - `indication`, `target` (и при желании класс препарата),
    - фактическое время разработки (в годах).
  - итог: модель **по паре (indication, target)** выдаёт **примерное время традиционной разработки** в годах.

- **`ai_time_years`**
  Получается из `traditional_time_years` по простой формуле:

  система моделирует ускорение за счёт ИИ, принимая, что
  AI-подход сокращает время примерно в 7 раз
  `ai_time_years = traditional_time_years / 7`.