Противоречие (Contradiction)

В анализе данных — ситуация, когда в двух записях множества данных одному и тому же набору значений входных атрибутов соответствуют различные наборы значений выходных.

Так, для задачи классификации это означает, что два объекта с одинаковыми признаками относятся к различным классам, что противоречит логике анализа.

Например, если два клиента банка имеют одинаковые параметры (доход, наличие недвижимости, возраст и т.д.), но при этом для одного из них кредитный рейтинг будет «Высокий», а для другого — «Низкий», то соответствующие примеры будут противоречивыми.

Возраст	Стаж, лет	Доход, руб.	Число иждивенцев	Наличие автомобиля	Наличие недвижимос
45	27	50 000	3	Да	Да
45	27	50 000	3	Да	Да

Наличие противоречий является одним из факторов, ухудшающих качество данных. Противоречия искажают закономерности в данных, поиск которых и является целью анализа, что приводит к снижению точности аналитических моделей. Поэтому исключение противоречий является одной из наиболее важных задач очистки данных.

При обработке противоречий возможны два подхода:

- Первый предполагает, что противоречие вызвано ошибкой (например, неправильно указана метка класса). В этом случае запись с ошибкой можно просто удалить.
- Второй подход допускает, что записи, хотя и являются противоречивыми, тем не менее отражают реальные события. В этом случае обычно производят объединение записей с агрегированием числовых значений выходных атрибутов.