Дубликат (Dublicate)

Синонимы: Копия
Решения: Loginom Data Quality

Две или более записи одного набора данных называются дубликатами, если они содержат идентичные наборы значений **всех** признаков.

В большинстве случаев дубликаты рассматриваются как негативный фактор, и в процессе очистки данных от них стремятся избавиться. Это связано с тем, что дублирующие записи (кроме одной) не несут никакой полезной информации и бесполезны с точки зрения обучения аналитических моделей.

Большое количество дубликатов обедняет обучающее множество в информационном плане: если из 1000 примеров обучающего множества 700 — дубликаты, то фактически для обучения используется всего 301 пример, хотя обрабатываются все.

Однако в некоторых случаях добавление дубликатов в обучающую выборку позволяет повысить эффективность обучения модели. Например, увеличивая или уменьшая число положительных или отрицательных примеров для бинарной классификационной модели путем дублирования, можно управлять соотношением ложноположительных и ложноотрицательных исходов классификации.