5 Azure ML Studio

5.1 Кластеризация

Для выполнения кластерного анализа данных используется блок `K-Means Clustering`.

Параметры, которые мы будем изменять:

- Number of Centroids – число центроидов.
- Initialization – первичная инициализация, будет использована инициализация на основе label-столбца с предварительно назначенными кластерами.
- Metric – метрика расстояния.
- Iterations – ограничение числа итераций.

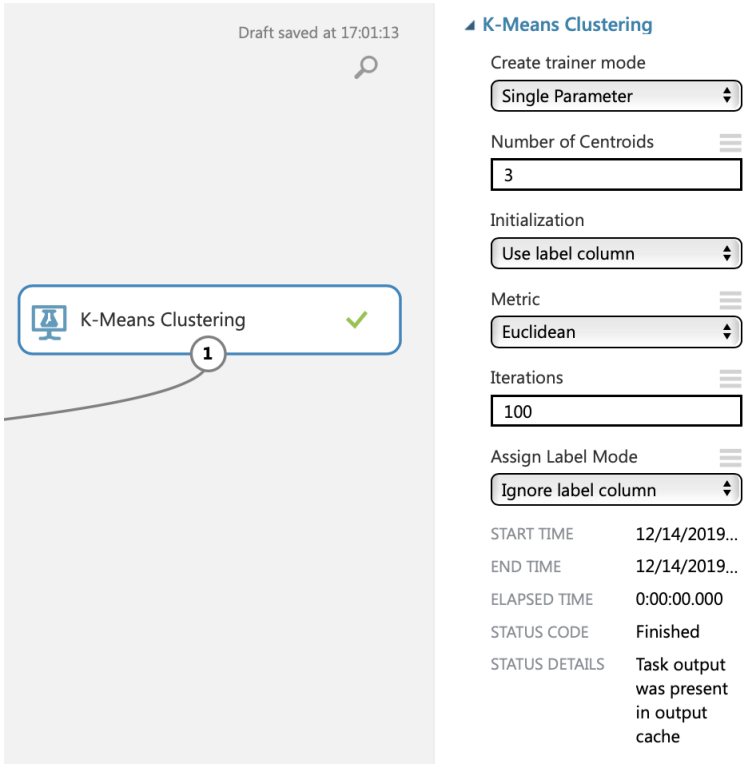

Рис. 1: Блок настроек кластерного анализа.

Блок `Train Clustering Model` отвечает за процесс обучения модели. На вход подаются данные и выбранный метод кластеризации. В параметрах данного блока необходимо выбрать столбцы предикторы.

В случае с инициализацией на основе label-столбца необходимо отметить столбец данных как label. Для этого следует настроить метаданные с помощью блока `Edit Metadata`, в котором необходимо выбрать колонку данных, отвечающую за инициализацию, а в поле Fields – значение Label.

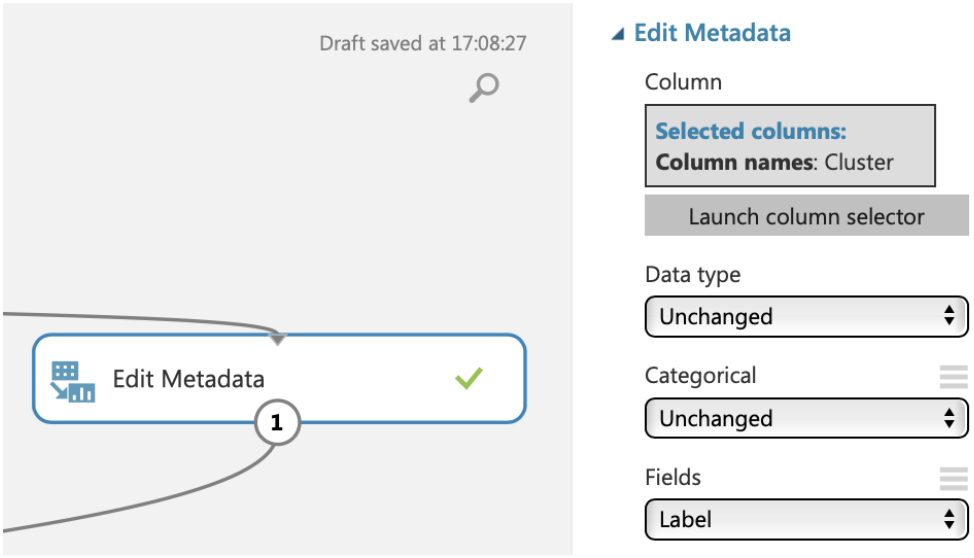

Рис. 2: Настройка метаданных.

После запуска модели, в блоке `Train Clustering Model` можно посмотреть визуализацию полученных кластеров. Для получения назначенных кластеров следует конвертировать данные в набор данных блоком `Convert to Dataset`.

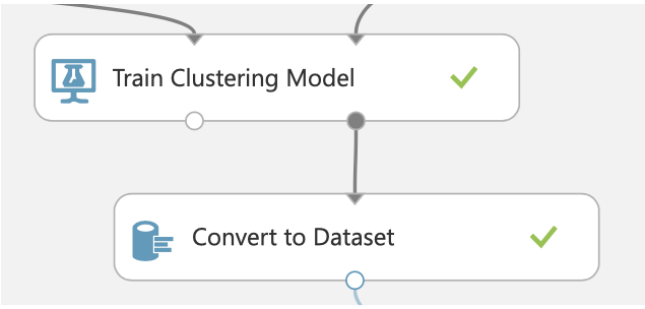

Рис. 3: Конвертация в набор данных.

5.2 Схема эксперимента

На рисунке ниже представлена общая схема модели кластеризации.

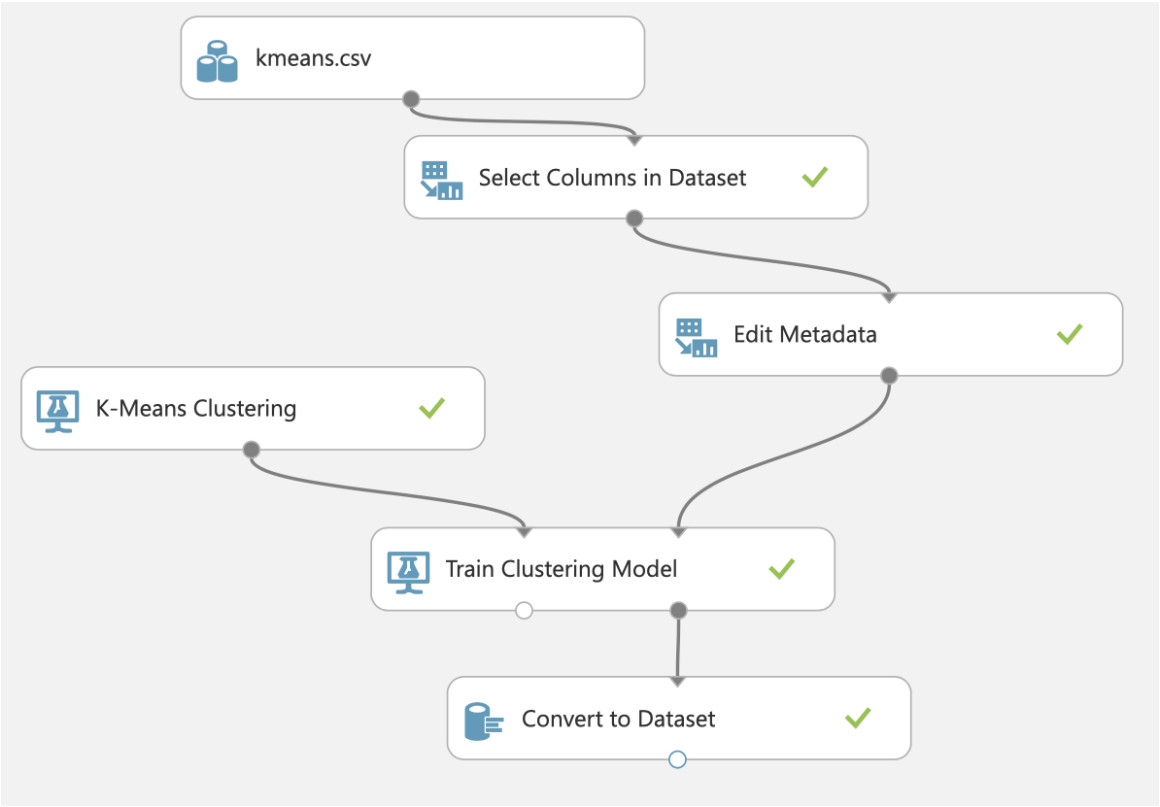

Рис. 4: Итоговый эксперимент.