1 Azure ML Studio

1.1 Данные

Исходные данные[1] – данные о 768 пациентах – женщины не менее 21 года индийского происхождения Пима.

Данные содержат следующие столбцы:

- Pregnancies – количество беременностей;
- Glucose – концентрация глюкозы в плазме крови через 2 часа при пероральном тесте на толерантность к глюкозе;
- Blood – диастолическое артериальное давление (мм рт. ст.);
- SkinThickness – толщина кожной складки трицепса (мм);
- Insulin – 2-часовой сывороточный инсулин (ед/мл);
- BMI – индекс массы тела ((вес в кг)/(квадрат роста в метрах));
- DiabetesPedigreeFunction – функция родословной (функция, которая оценивает вероятность развития диабета на основе семейного анамнеза);
- Age – возраст (лет);
- Outcome – переменная класса (0 или 1). 268 из 768 – это 1 (больны диабетом), остальные – 0.

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database

1.2 Разделение данных

Воспользуемся лишь частью данных и выберем первые 700 строк с помощью SQL запроса.

Разделим данные в соотношении 80/20 неслучайным образом (первые 80 процентов строк окажутся в первой группе, оставшиеся 20 — во второй). Для этого снимем галочку `Randomized split` блока `Split Data`. Параметр `Fraction of rows in the first output dataset` зададим равным `0.8`, тогда на первом выходе получим данные для обучения модели, на втором выходе – для ее оценки. Значение `Random seed` можно не указывать, так как случайное разделение не используется.

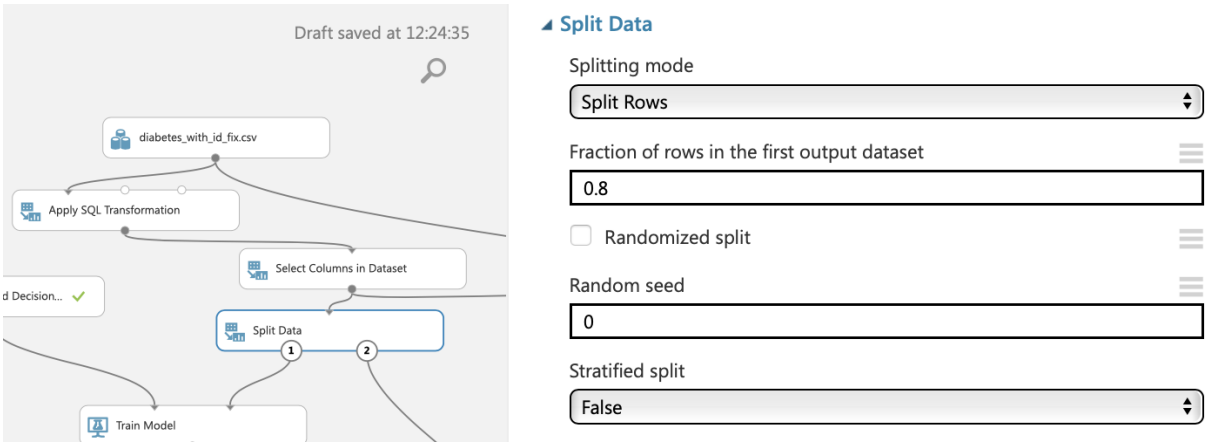

Рис. 1: Разделение данных.

1.3 Модель классификатора

Для обучения модели дерева принятия решений используется блок `Two-Class Boosted Decision Tree` из раздела `Machine Learning`. Интересующие нас параметры:

- `Number of trees constructed` – число создаваемых деревьев. Так как мы изучали только построение одного ДПР, то присваиваем этому полю значение 1. Большее число деревьев нам понадобится при изучении случайного леса в следующих лекциях;
- `Maximum number of leaves per tree` – максимальное числов листьев в дереве;
- `Minimum number of samples per leaf node` – минимальное число объектов в листе;
- `Random number seed`.

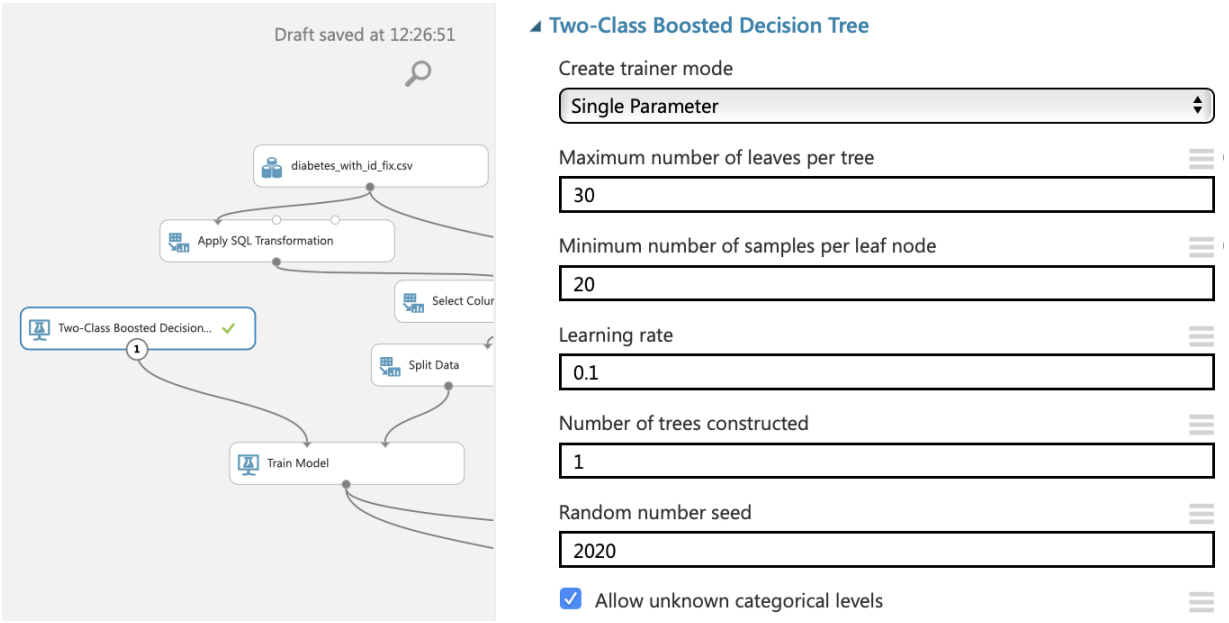

Рис. 2: Параметры блока `Two-Class Boosted Decision Tree`.

Блок `Train Model` все также отвечает за обучение модели. На вход подаются данные и выбранный метод машинного обучения. В качестве предикторов выступают колонки `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction`, `Age` и отклик `Outcome`. В параметрах данного блока необходимо выбрать столбец данных, соответствующий отклику.

После запуска модели, в параметрах блока `Train Model`, пункт `Visualize` можно посмотреть построенное дерево и все критерии. Рекомендуем поставить галочку в пункте `Edges`.

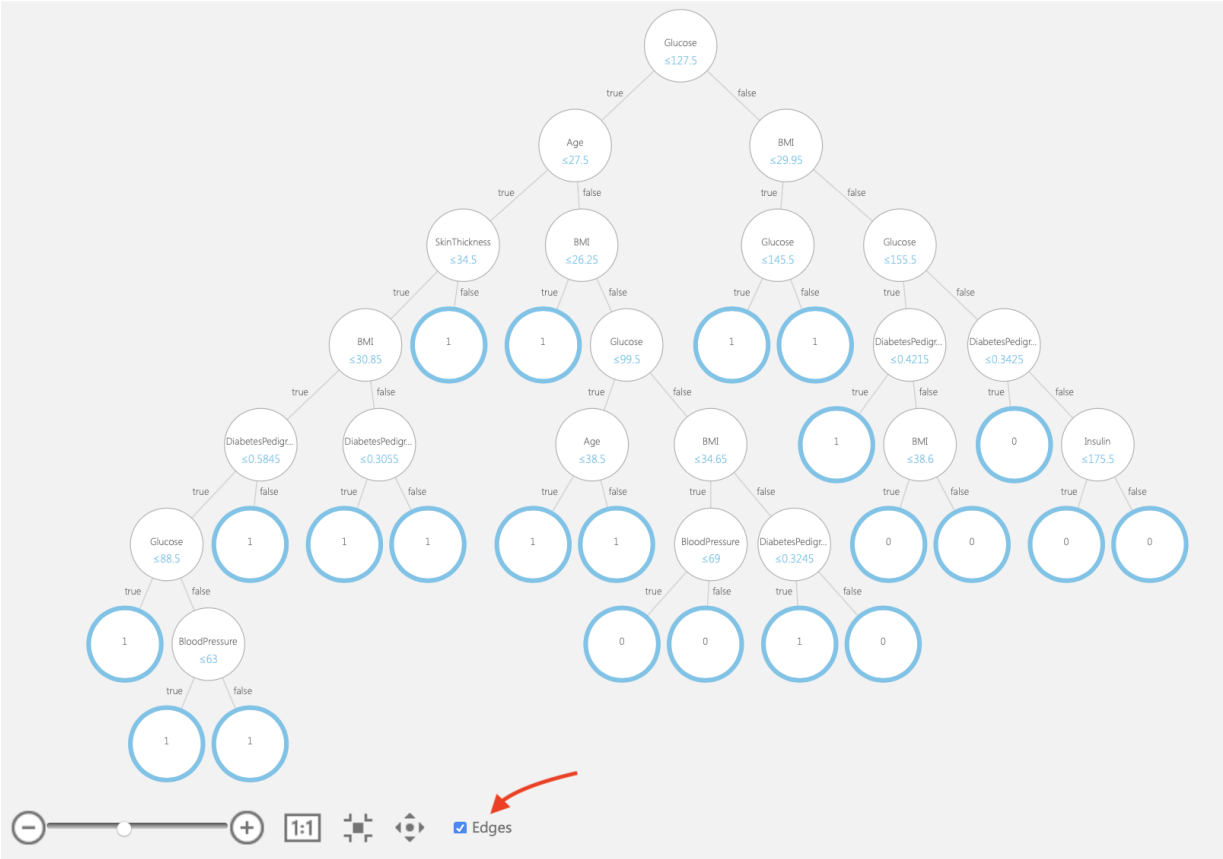

Рис. 3: Параметры блока `Two-Class Boosted Decision Tree`.

1.4 Задача классификации

Оценка модели и предсказание для новых пациентов, с помощью обученной модели выполняются аналогичным любым классификаторам образом.