Метод «k ближайших соседей»

1. Чтобы сделать прогноз для новой точки данных, алгоритм находит ближайшие к ней обучающие точки набора («ближайшие соседи») и выбирает доминирующий среди них класс.
2. В методе определён параметр n\_neighbors, задающий количество ближайших соседей, используемых для предсказания. Кроме числа соседей можно также задать веса (параметр weights) отдельных точек. Если выбрать в качестве весов расстояние (weights = ‘distance’), то для определения класса точки будет использоваться не только число соседей, но и расстояние от неё до них. Также в методе определён параметр metrics, задающий функцию для измерения расстояния.
3. Масштабирование признаков необходимо для того, чтобы сделать набор данных менее разрежённым и повысить точность метода.
4. Недостатки:
   1. Алгоритм работает хуже при увеличении количества признаков или когда у большей части признаков нулевые значения.
   2. Всегда необходимо определять оптимальное количество соседей.

Преимущества:

1. Лёгкая интерпретация.
2. Малое количество настроек.
3. Быстрое построение модели.
4. Алгоритм не чувствителен к выбросам.
5. При кросс-валидации вместо простого разбиения данных на обучающую и тестовую выборки происходит разбиение на несколько частей, и каждая из частей по очереди выступает в качестве тестовой, а все остальные в качестве обучающей. Качество модели оценивается на основе среднего значения качества для всех таких разбиений. Этот метод даёт более точную оценку качества модели, чем простое разбиение выборки на обучающую и тестовую.