Презентация по курсовой работе на тему:

«Исследование метрических методов классификации»

Выполнил:
Ощепков Дмитрий
Олегович

Научный руководитель:
Котельников Евгений
Вячеславович

Актуальность

Метрические методы классификации применяются в различных областях жизнедеятельности человека, например:

- Биомедицина и медицинская диагностика
- Обработка естественного языка
- Компьютерное зрение и обработка изображений

Проблема исследования

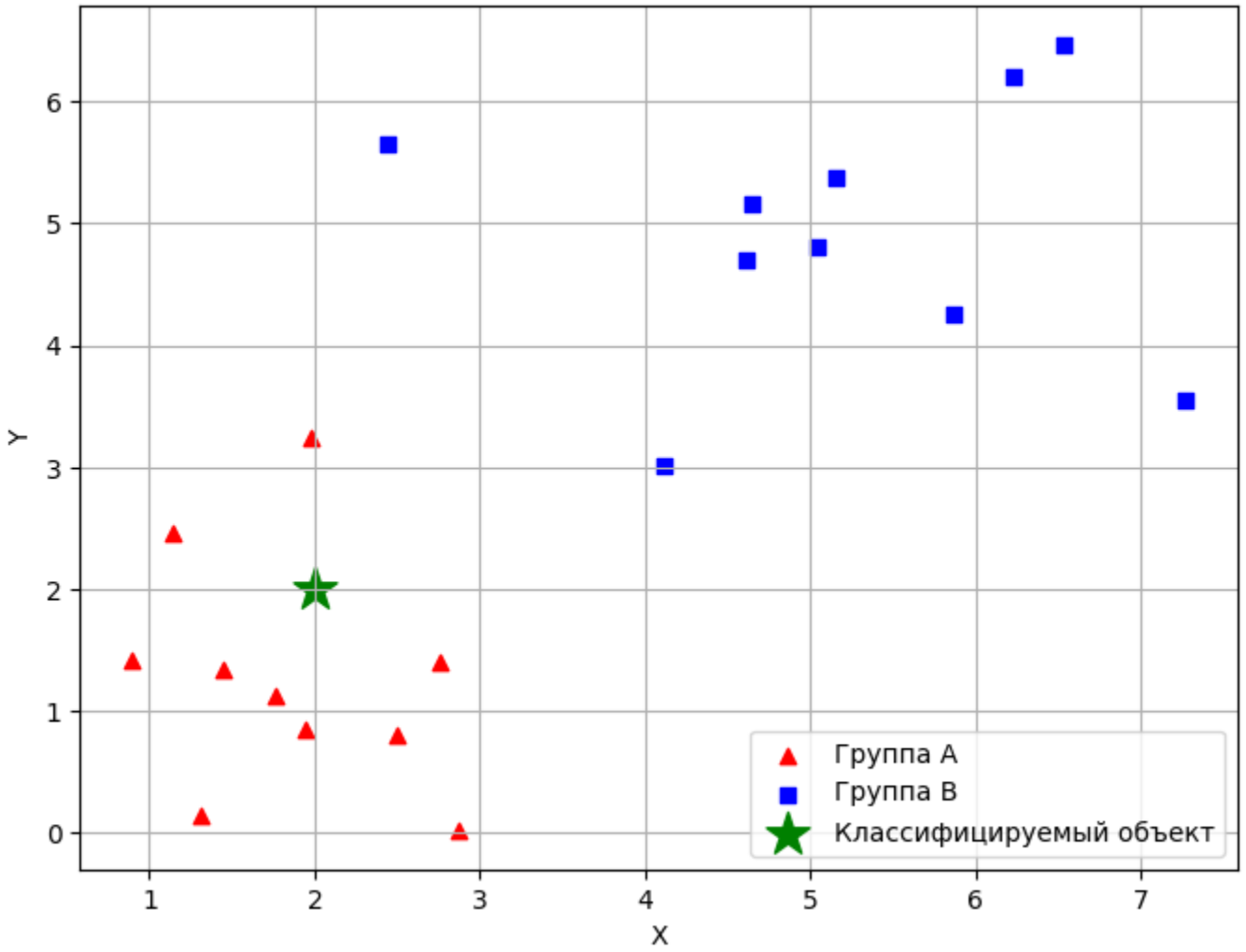
- Проблема исследования заключается в отсутствии четких рекомендаций или наставлений для выбора конкретного метрического метода классификации в конкретной задаче.

Цели и задачи

- Провести сравнительный анализ различных метрических методов классификации в разных сценариях и задачах.
- Сформировать рекомендации по выбору модели.

Постановка задачи

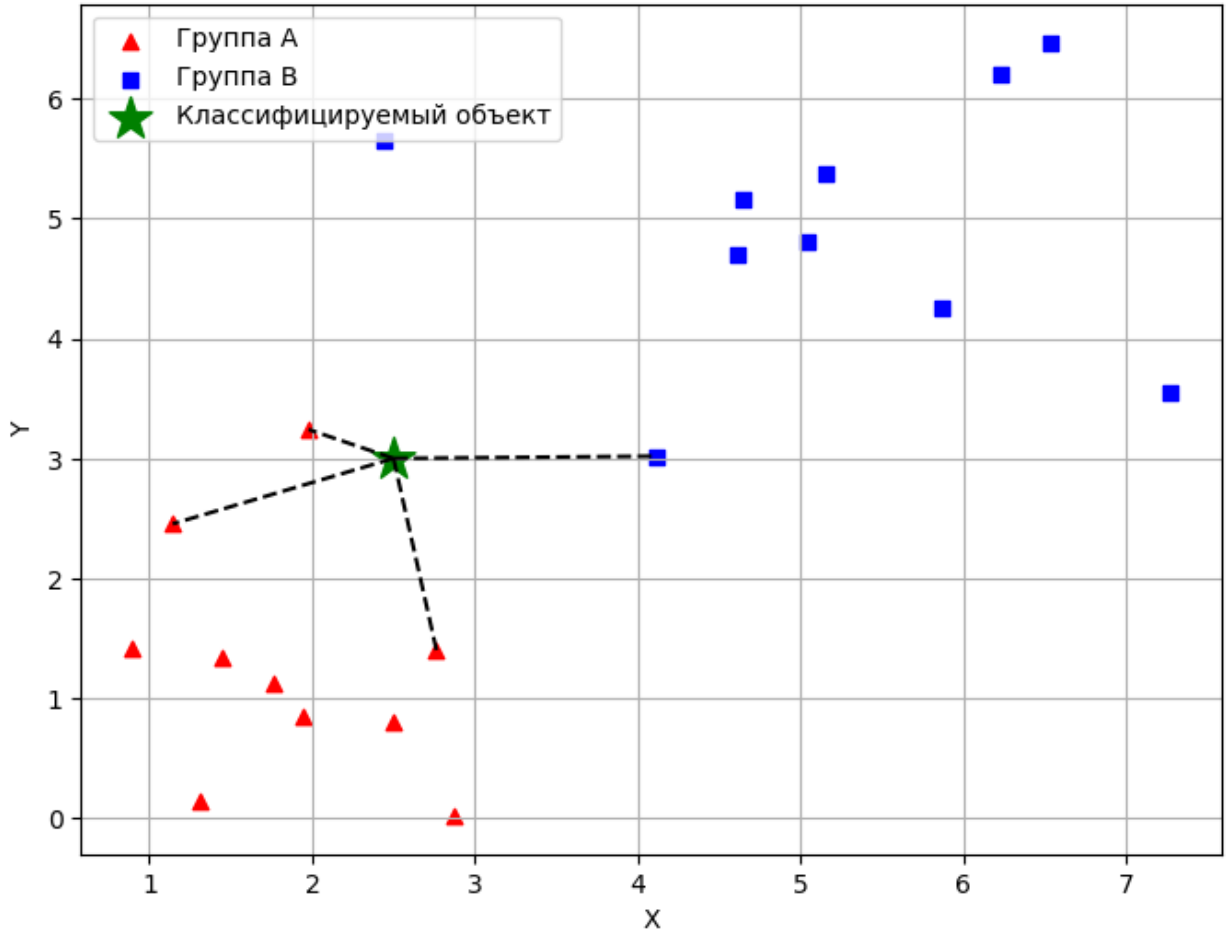
- Задача классификации заключается в том, чтобы разделить разные ситуации или объекты на группы или классы на основе имеющихся данных.



Метод k ближайших соседей

- Классификация на основе меток k ближайших объектов

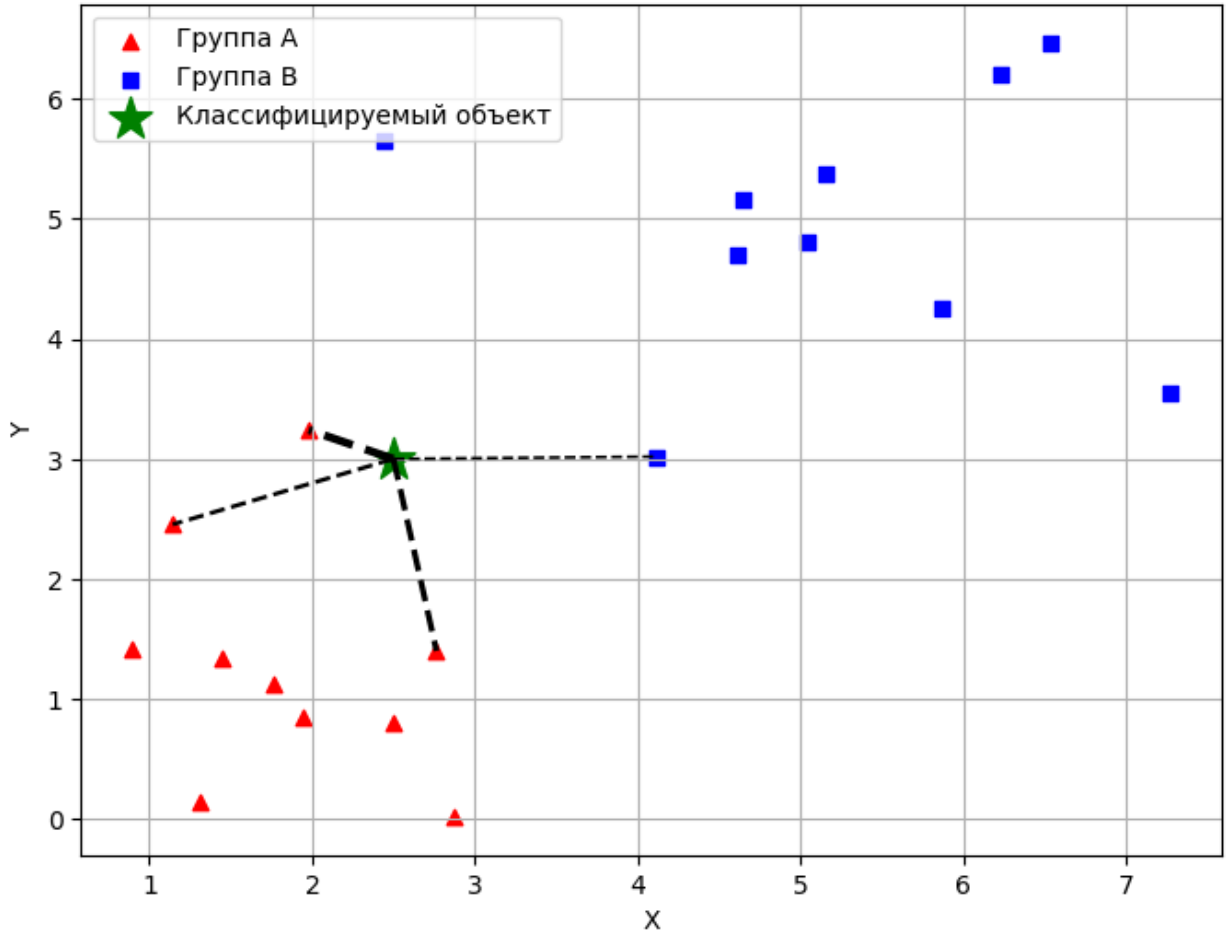
$$a(u, \underset{k}{k}) = \\ = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1} [y_u^{(i)} = y]$$



Метод взвешенных ближайших соседей

- Модификация предыдущего, при классификации присваивает веса ближайшим соседям

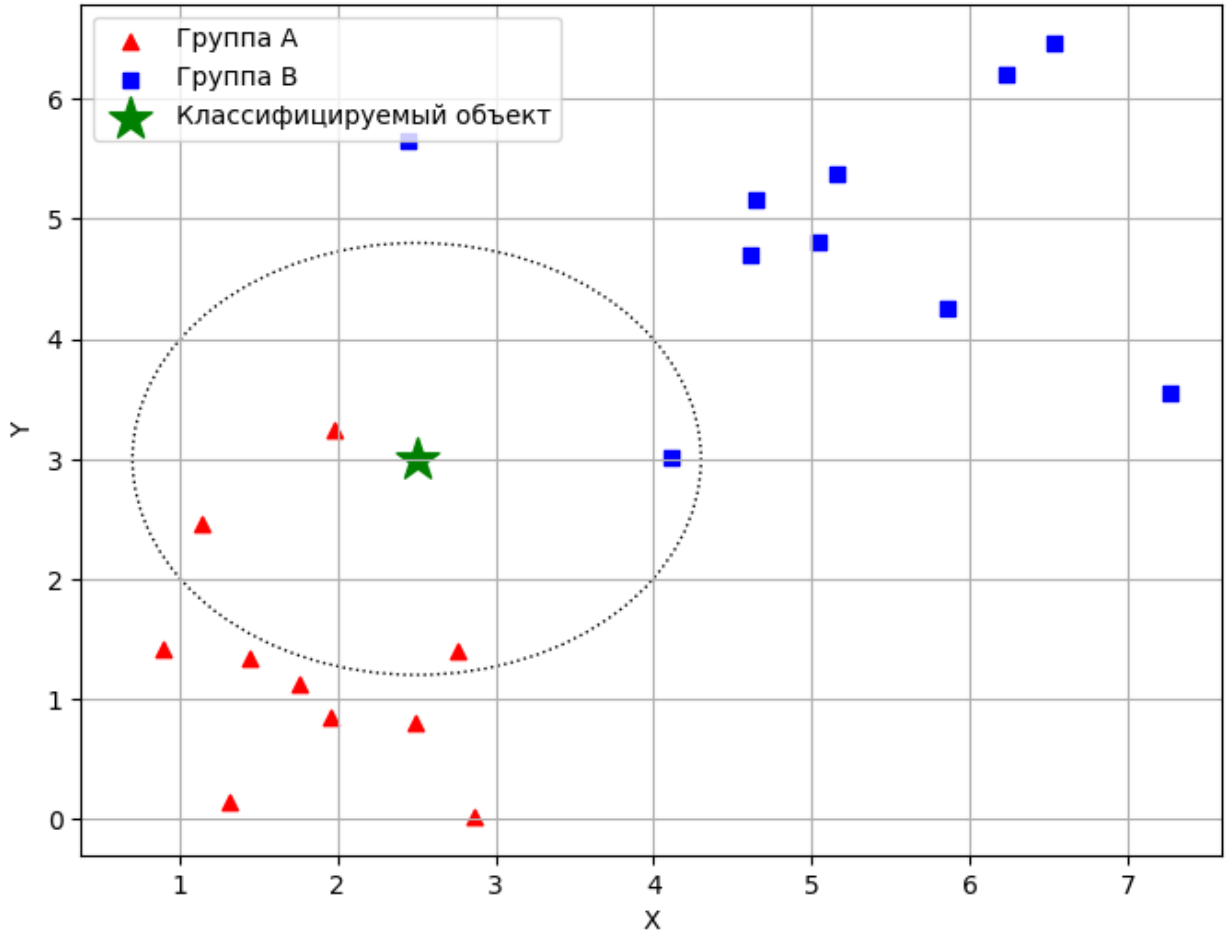
$$a(u, k) = \underset{l}{} = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1} \left[y_u^{(i)} = y \right] \mathrm{w_i}$$



Метод парзеновского окна фиксированной ширины

- В этом методе объект классифицируется на основе ядерной функции. Ширина окна влияет на важность удаленности объектов при классификации

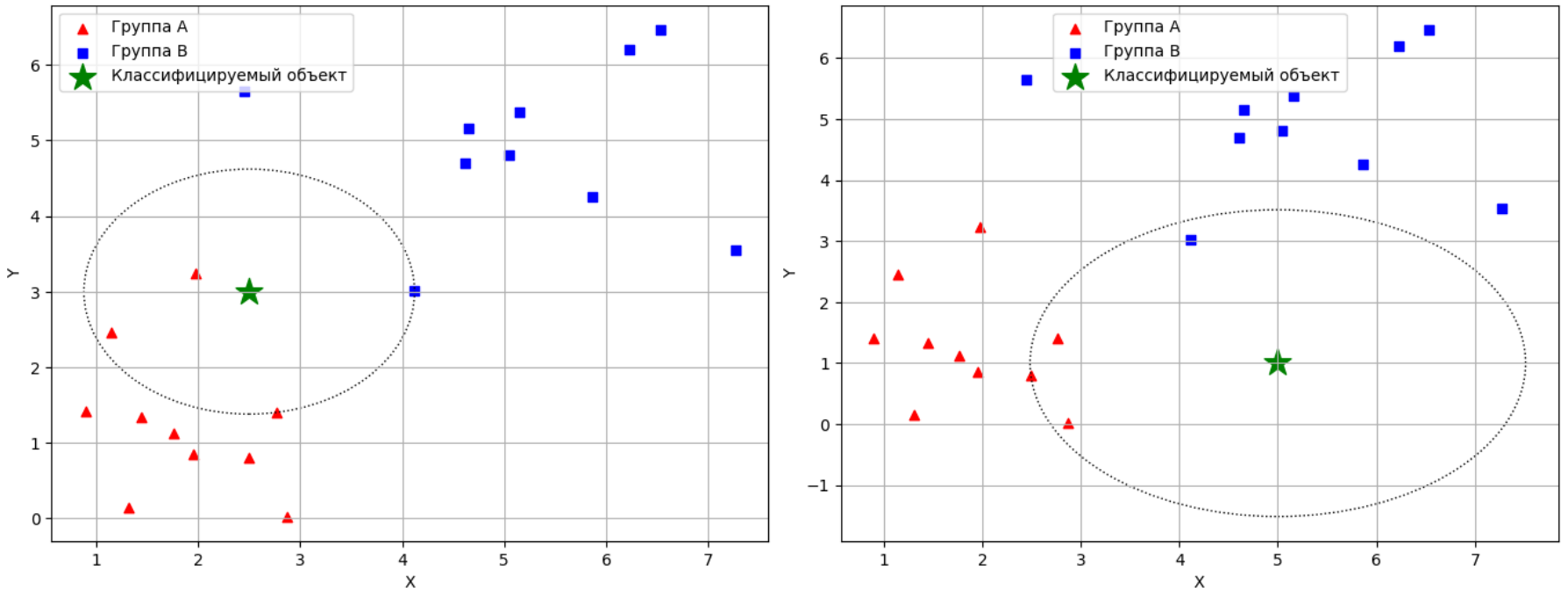
$$a(u, \mathrm{h}, K) = \\ = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1}^{l} \left[y_u^{(i)} = y \right] K\left(\frac{\rho\left(u, x_u^{(i)}\right)}{h} \right)$$



Метод парзеновсего окна переменной ширины

- В отличие от метода парзеновского окна фиксированной ширины, здесь ширина окна может изменяться в зависимости от расстояния между объектами

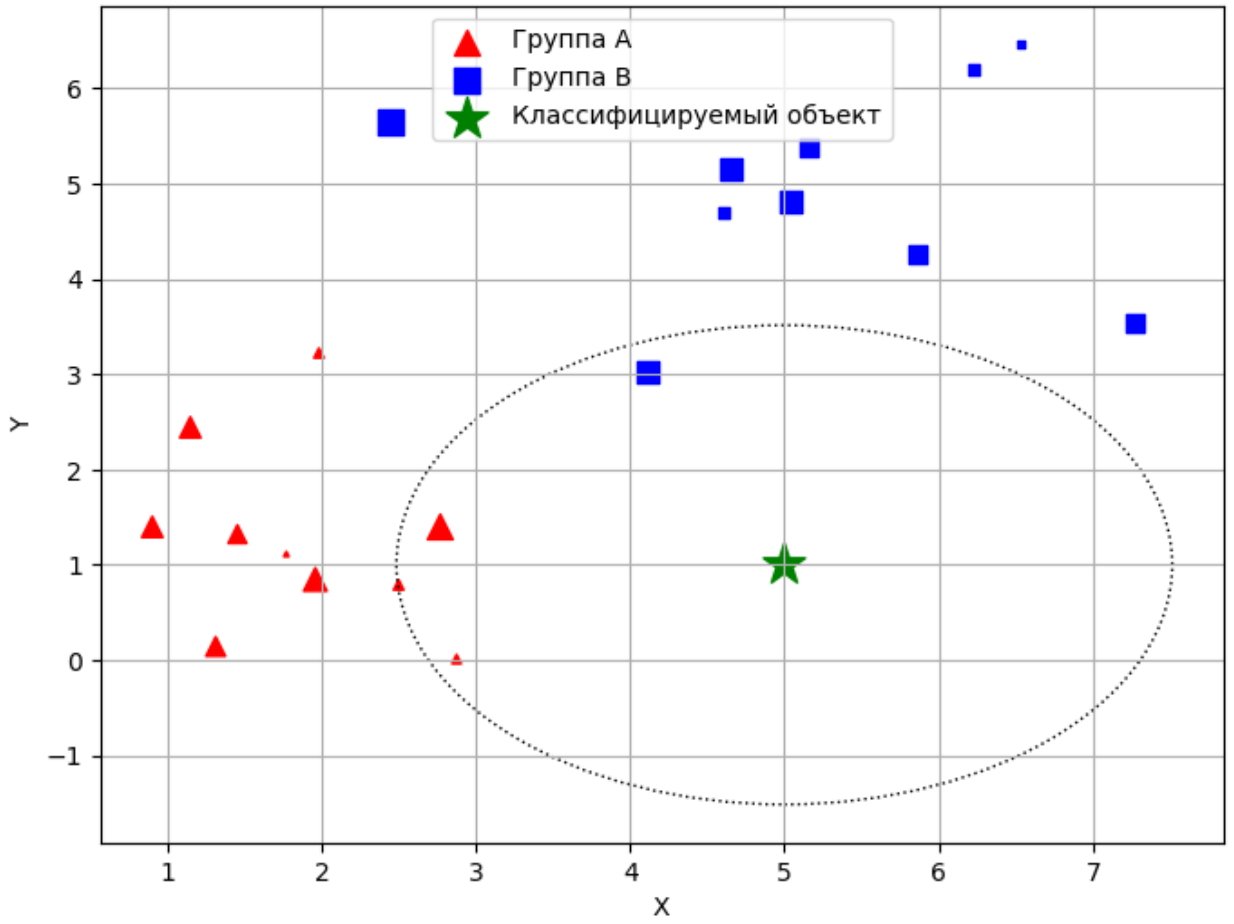
$$a(u, \mathrm{k}, K) = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1}^{k} \left[y_u^{(i)} = y \right] K \left(\frac{\rho\left(u, x_u^{(i)}\right)}{\rho\left(u, x_u^{(k+1)}\right)} \right)$$



Метод потенциальных функций

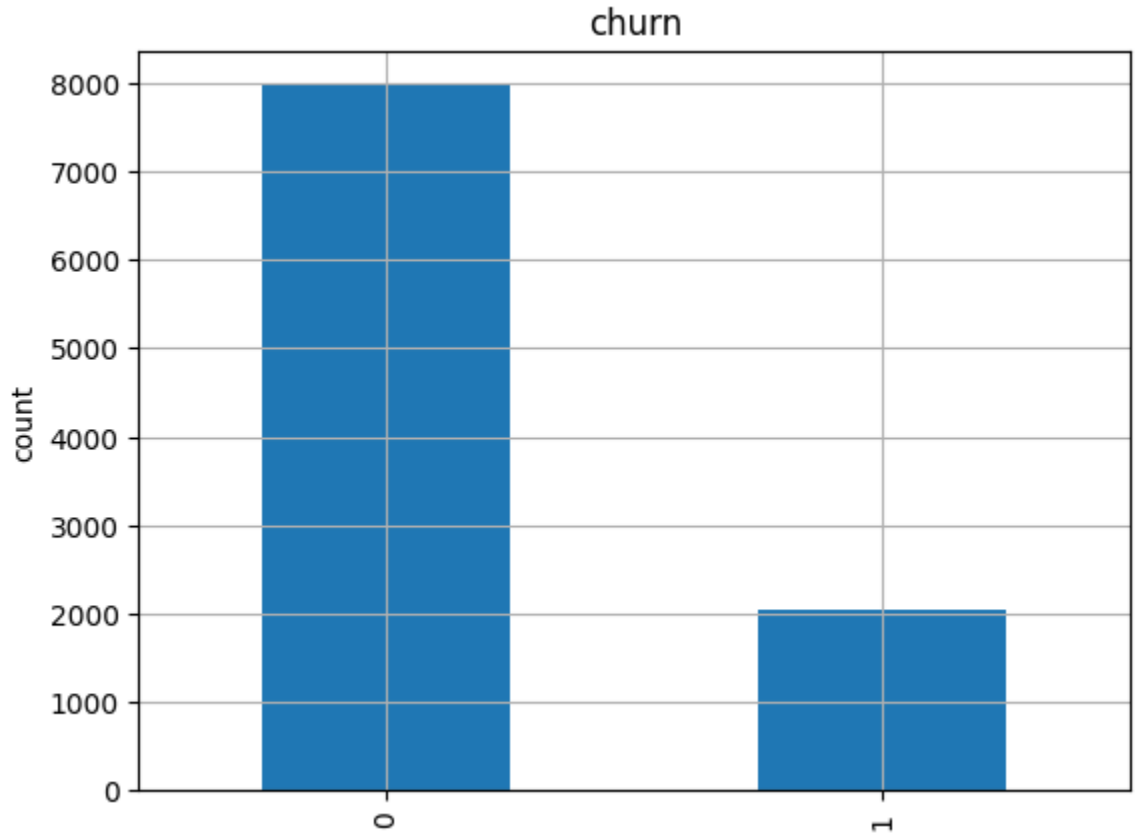
- Функция потенциала определяет близость объектов и их принадлежность к классам

$$a(u, K) = \underset{k}{} = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1} \left[y_u^{(i)} = y \right] * K \left(\frac{\rho\left(u, x_u^{(i)}\right)}{\rho\left(u, x_u^{(k+1)}\right)} \right)$$



Bank Customer Churn Dataset

- Из особенностей датасета можно выделить дисбаланс данных и отсутствие выбросов.
- Дисбаланс был устранен с помощью алгоритма SMOTE, который добавил примеры в меньший класс



Bank Customer Churn Dataset

	KNN	Weight KNN	Fixed-width Parzen Window	Variable-width Parzen Window	Potential Function	Sk-learn
Accuracy	0.7753	0.7753	0.7093	0.744	**0.798**	0.7753
F1	**0.4871**	**0.4871**	0.4591	0.4754	0.0066	**0.4871**
Precision	**0.4482**	**0.4482**	0.3656	0.4028	0.2	**0.4482**
Recall	0.5333	0.5333	**0.6167**	0.58	0.0033	0.5333

Bank Customer Churn Dataset

- Были проведены замеры метирик до и после применения SMOTE
- На практике подтвердилось, что рассмотренные методы чувствительны к дисбалансу классов.
- Метод потенциальных функций показал наибольшее accuracy, но базовый классификатор выдает 0.7753, что говорит о низком уровне качества модели
- Наилучшие показатели основной метрики (F1) у knn и knn-weighed
- Стоит отметить, что Fixed-width Parzen Window имеет наибольший показатель Recall

Остальные датасеты

- ….

Выводы

- В зависимости от задачи могут быть применимы разные методы, но наибольшие показатели в тестах как правило у алгоритма knn, поэтому этот алгоритм стоит пробовать первым.
- Алгоритмы действительно чувствительны к нормировке, (выбросам) и дисбалансу классов

Спасибо за внимание!