Разработка механизма определения аномалий в данных

Студент 4 курса, 592 гр.:
Каразеев А. А.

Научный руководитель:
Дайняк А. Б.

МФТИ, 2019

Цель работы

- **Объектом исследования** являются методы для экстраполяции временных рядов и поиска выбросов в данных, а также способы их реализации и области применения.

- **Предметом исследования** является анализ существующих алгоритмов для обработки данных.

- **Методы исследования** — анализ предметной области, анализ реализованных методов, написание программного кода и извлечение полезной информации из данных.

Поставленные задачи

- Изучение уже созданных алгоритмов для поиска аномалий в данных. Поиск датасетов.

- Анализ эффективности алгоритмов на релевантных данных.

- Программная реализация алгоритмов с использованием языка программирования Python.

- Встраивание алгоритмов в сервис для обработки данных.

Рассматриваемые алгоритмы

- **k-Nearest Neighbors** (k-NN) – метод k-ближайших соседей.
- **Principal Component Analysis** (PCA) – метод главных компонент.
- **One-Class Support Vector Machines** (OCSVM) – одноклассовый метод опорных векторов.
- **Local Outlier Factor** (LOF) – метод локального уровеня выброса.
- **Histogram-Based Outlier Score** (HBOS) – оценка выбросов на основе гистограммы.
- **Isolation Forest** (IFOREST) – метод изолируещего леса.

Используемые датасеты

- **Arrhythmia** – определение наличия аритмии по данным ЭКГ.
- **Breast Cancer** – определение типа опухоли молочной железы: доброкачественная или злокачественная.
- **Glass** – идентификация типа стекла, оставленного на месте преступления.
- **Ionosphere** – рассматриваются характеристики радаров, которые используется в анализе ионосферы: необходимо определить является радар ”плохим”или ”хорошим”.
- **Letter Recognition** – по описанию изображения определить присутствует ли буква из английского алфавита или нет.
- **Mammography** – детектирование микрокальцинатов по данным маммографии.
- **MNIST** – научиться различать изображения рукописных цифр 6 и 0.
- **Satellite** – определение типа почвы по спутниковым снимкам.

Источник данных: Outlier Detection DataSets (ODDS) — http://odds.cs.stonybrook.edu

Используемые датасеты

Таблица 1 — Статистика по данным из рассматриваемых датасетов.

Датасет	Кол-во объектов	Размерность	Процент выбросов
arrhythmia	452	274	14.60
breastw	683	9	34.99
glass	214	9	4.21
ionosphere	351	33	35.90
letter	1600	32	6.25
mammography	11183	6	2.33
mnist	7603	100	9.21
satellite	6435	36	31.64

Код доступен по адресу: https://github.com/akarazeev/Bachelor

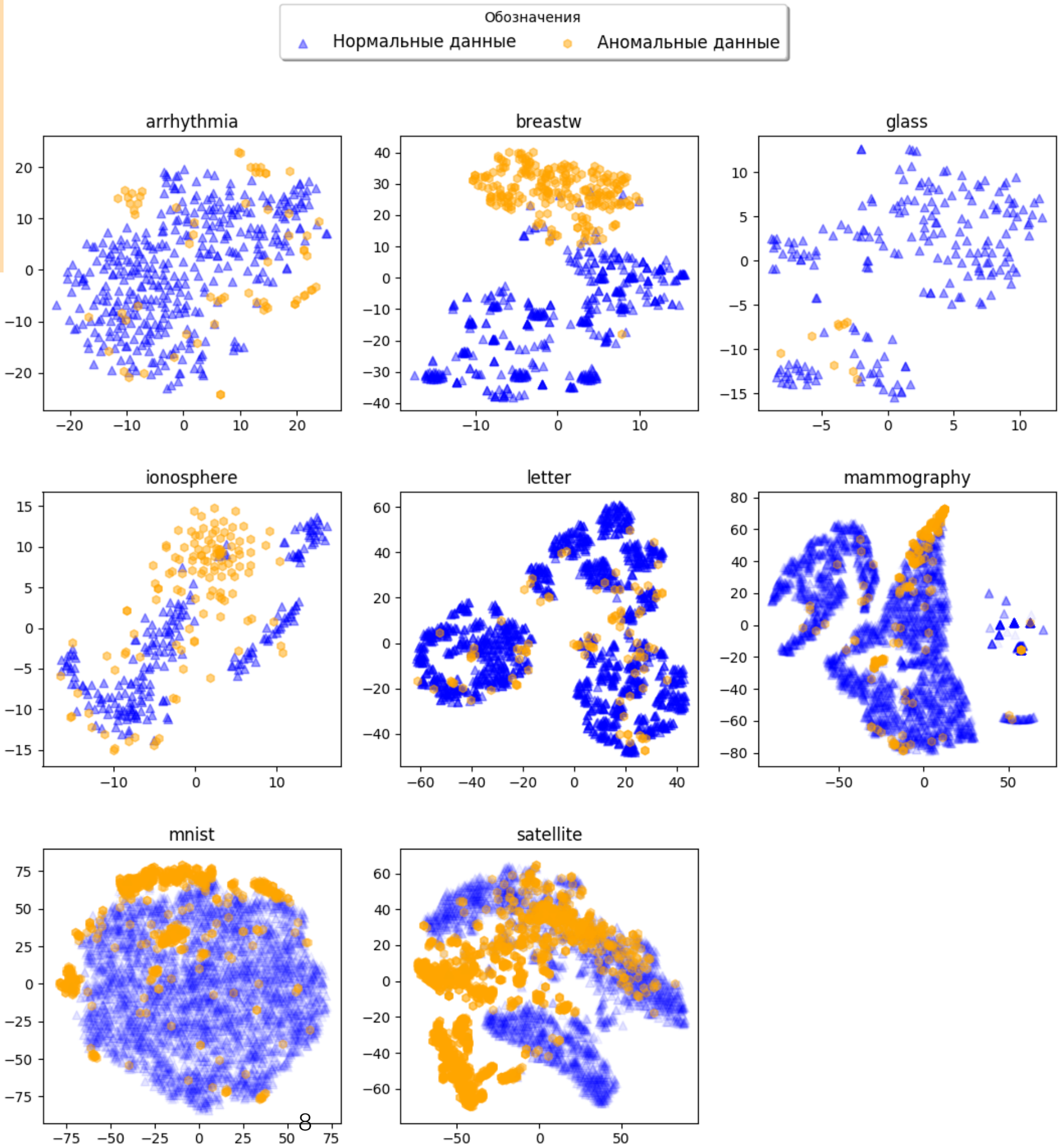
Используемые датасеты

Таблица 2 — Значения ROC для рассматриваемых алгоритмов на данных.

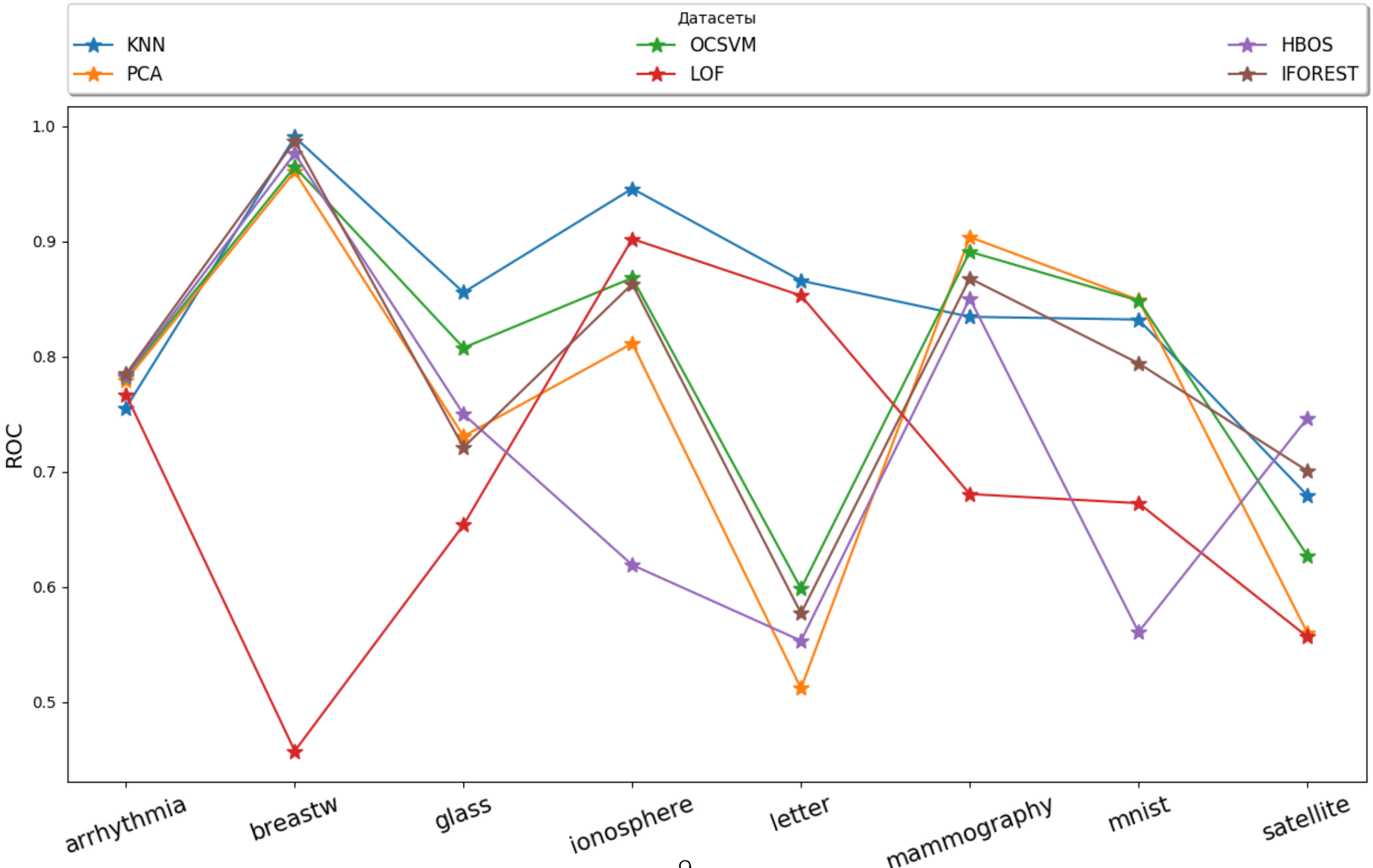
Датасет	KNN	PCA	OCSVM	LOF	HBOS	IFOREST
arrhythmia	0.7555	0.7794	0.7825	0.7672	0.7831	**0.7849**
breastw	**0.9908**	0.9608	0.9649	0.4574	0.9764	0.9872
glass	**0.8558**	0.7308	0.8077	0.6538	0.7500	0.7212
ionosphere	**0.9460**	0.8115	0.8684	0.9023	0.6190	0.8632
letter	**0.8660**	0.5119	0.5985	0.8530	0.5532	0.5770
mammography	0.8346	**0.9039**	0.8911	0.6806	0.8506	0.8680
mnist	0.8322	**0.8493**	0.8487	0.6727	0.5607	0.7942
satellite	0.6795	0.5601	0.6274	0.5567	**0.7464**	0.7008

Код доступен по адресу: https://github.com/akarazeev/Bachelor

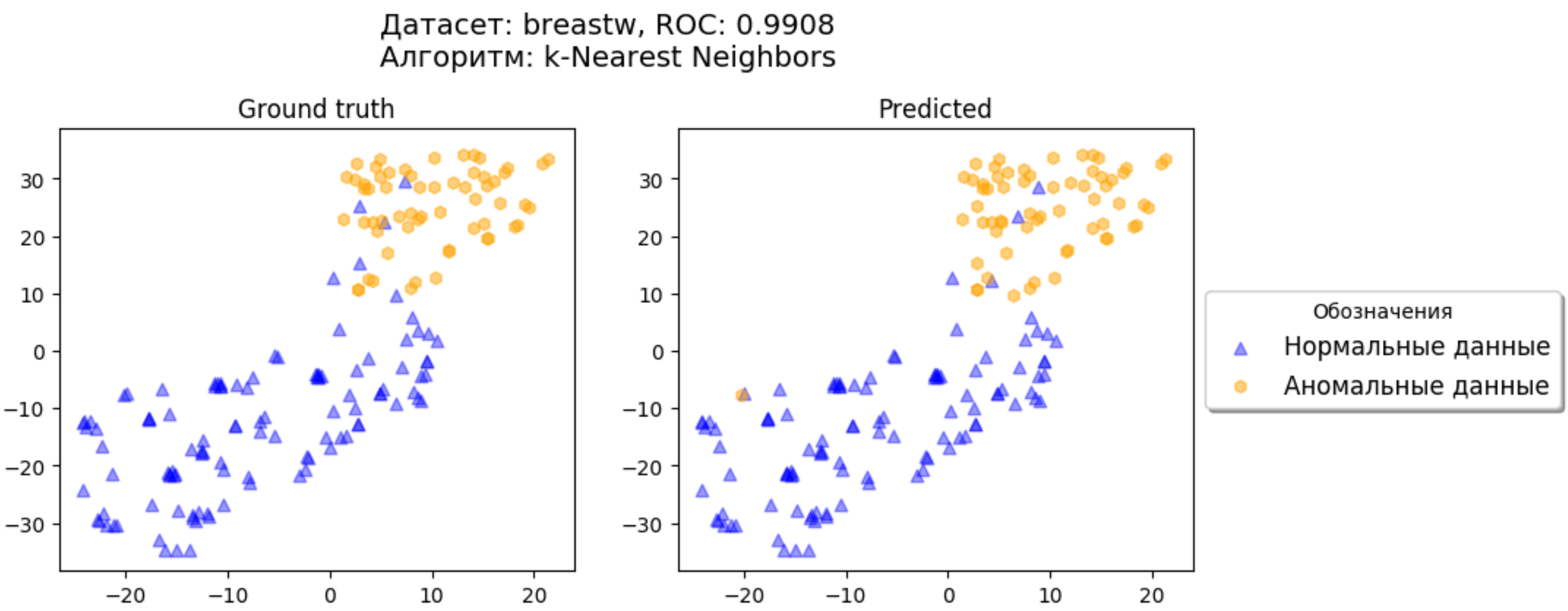
Представление объектов из датасетов после понижения размерности с помощью алгоритма t-SNE



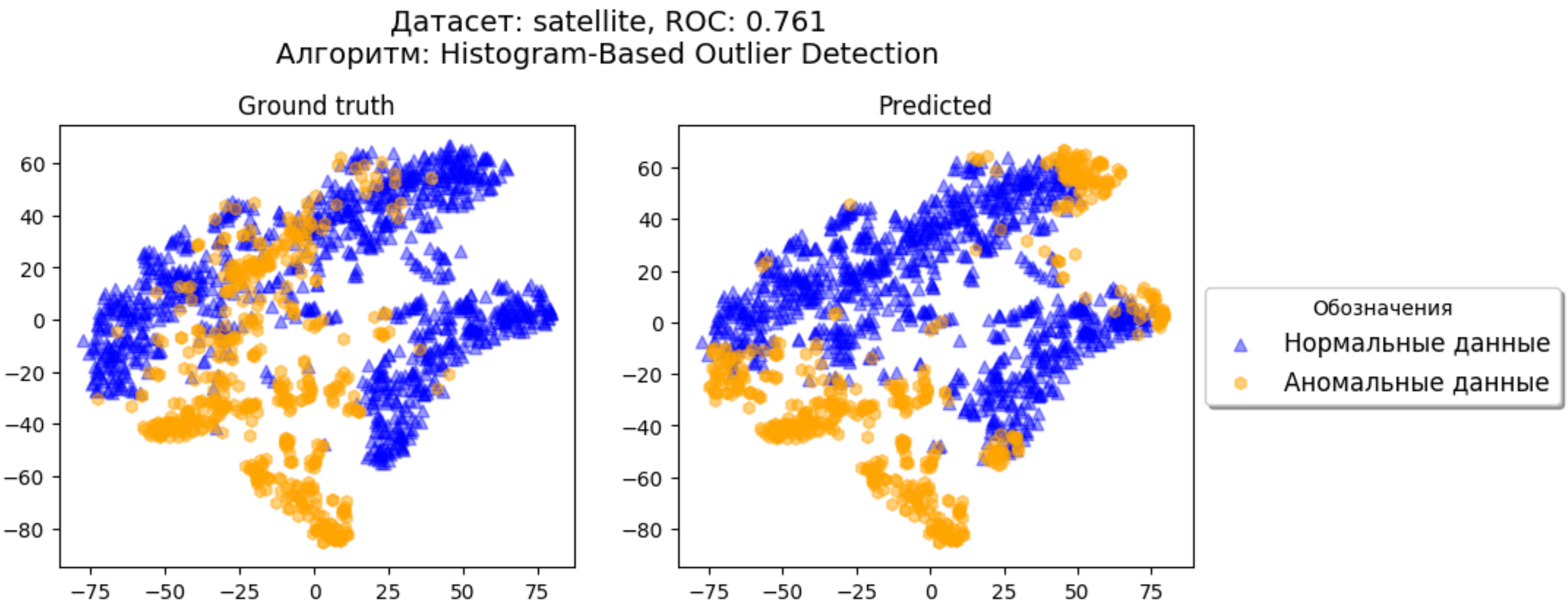
Результаты обучения алгоритмов



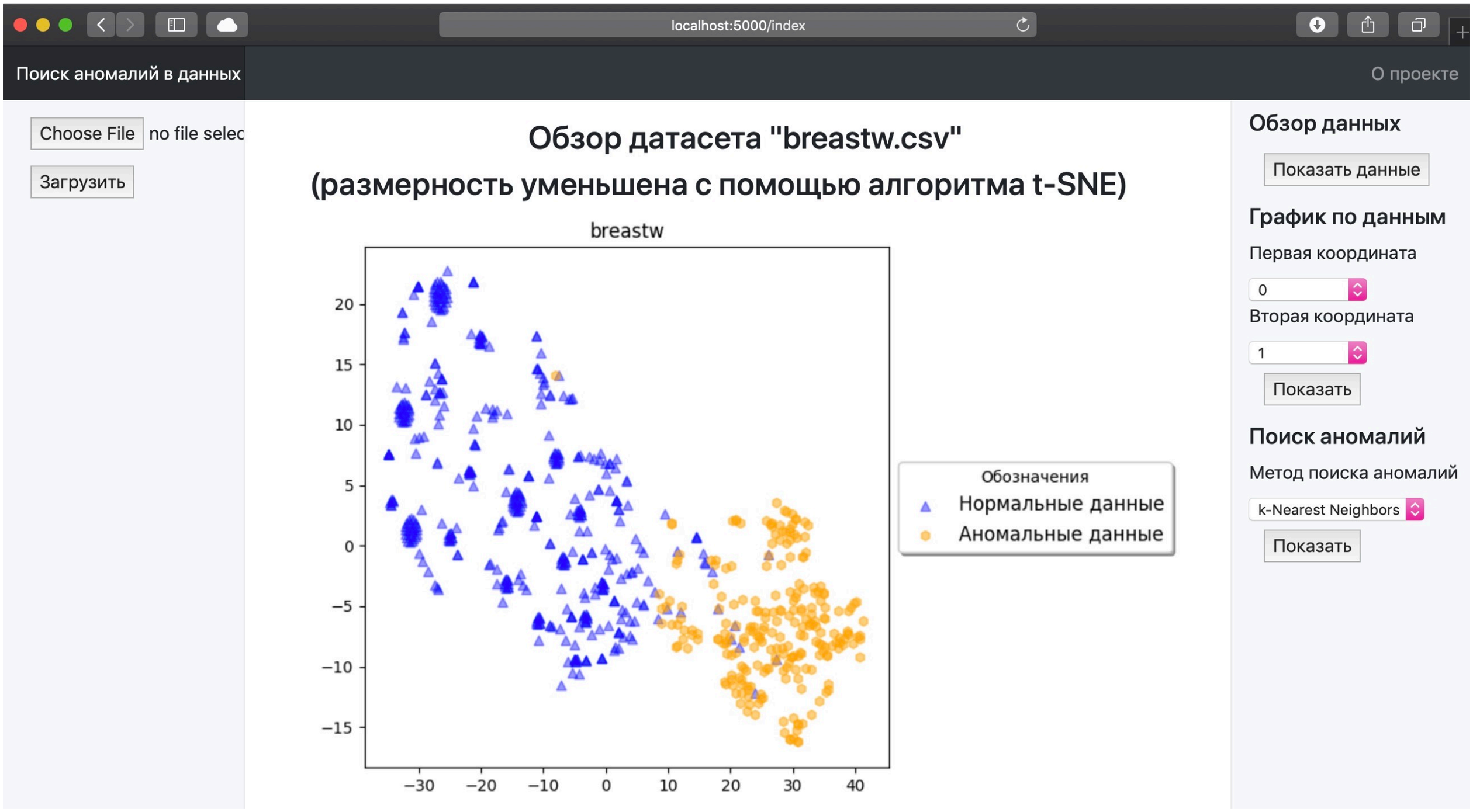
Результаты обучения алгоритмов



Результаты обучения алгоритмов

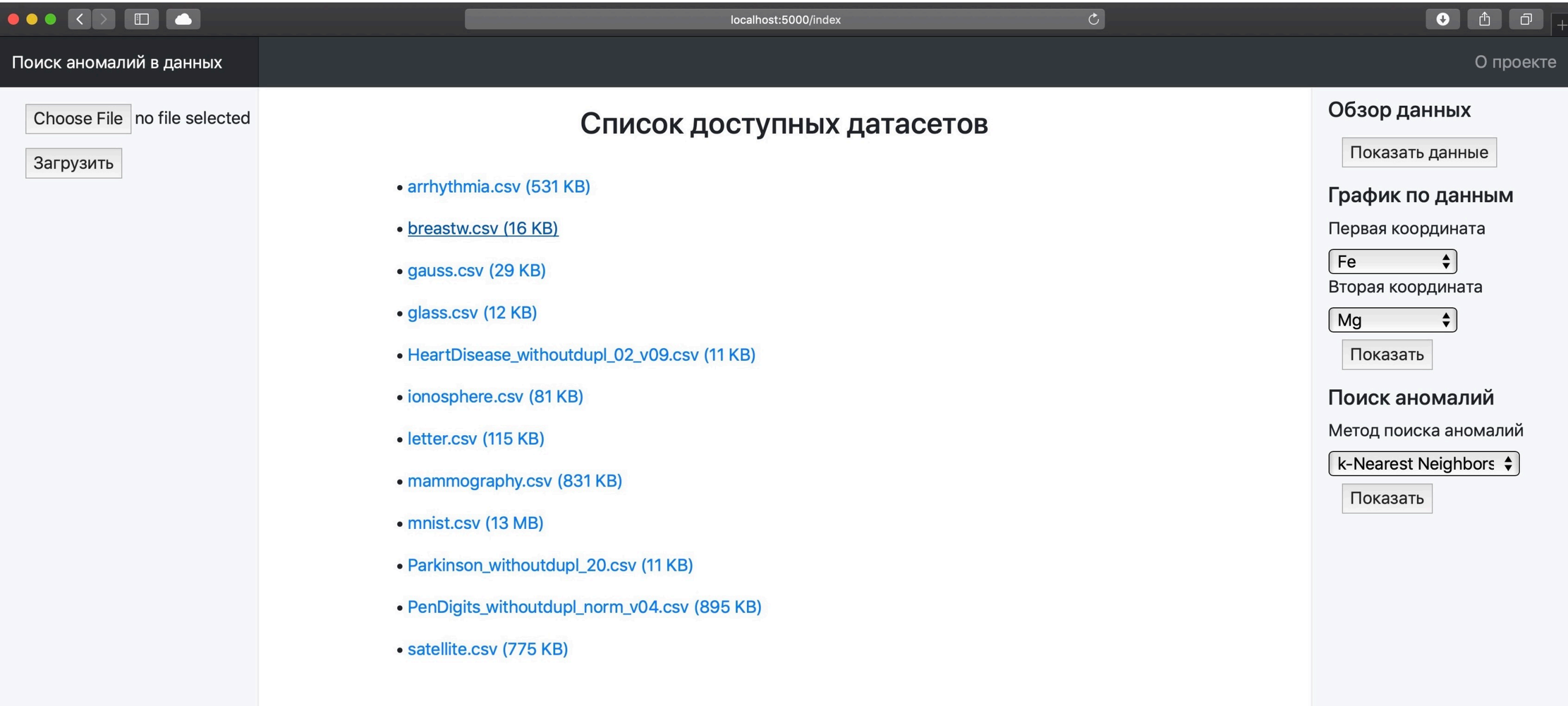


Сервис



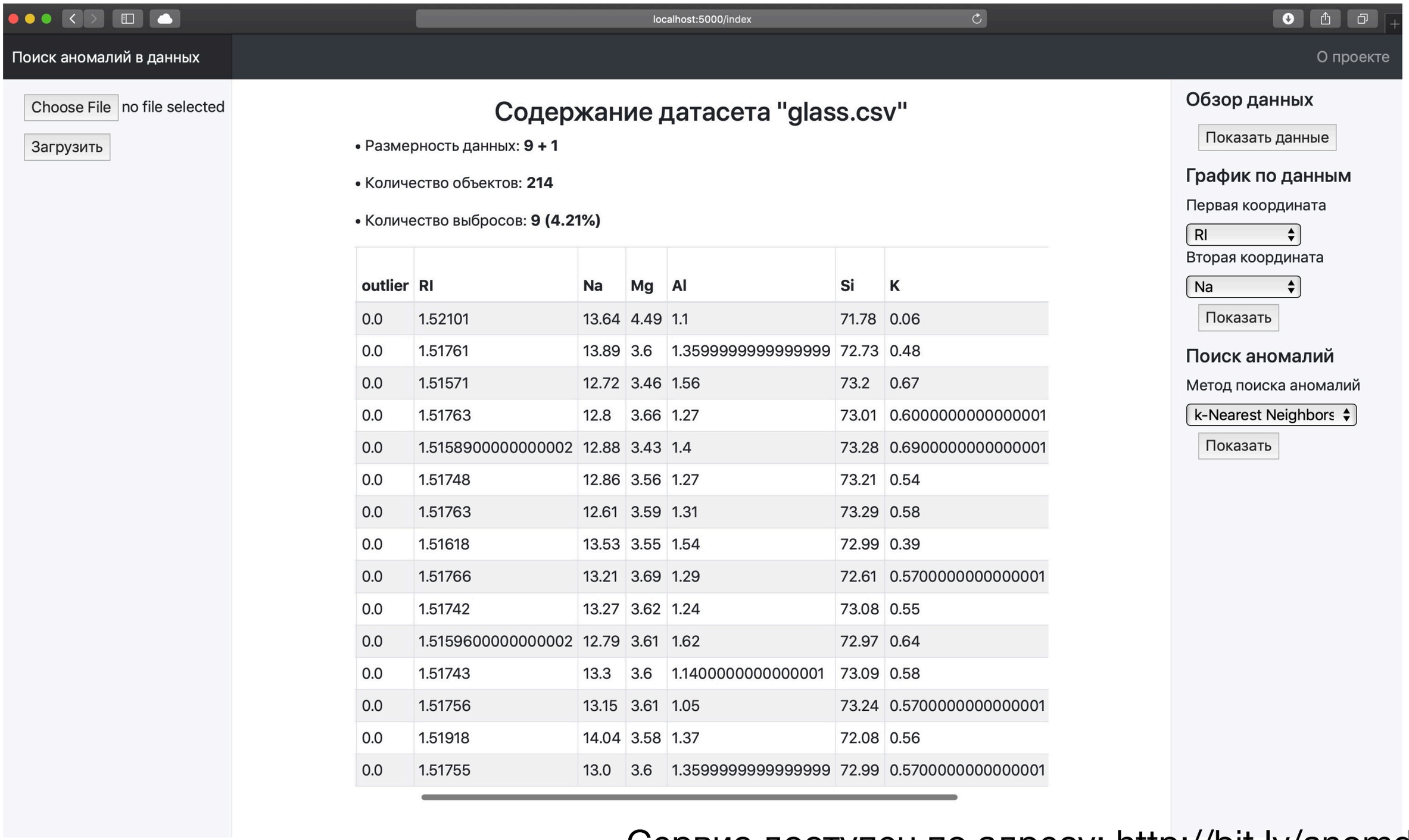
Сервис доступен по адресу: http://bit.ly/anomd19

Сервис



Сервис доступен по адресу: http://bit.ly/anomd19

Сервис



Сервис доступен по адресу: http://bit.ly/anomd19

Результаты проделанной работы

- Проанализированы актуальные алгоритмы для поиска аномалий в данных.
- Продемонстрированы релевантные данные и даны оценки качества рассмотренных алгоритмов на них.
- Предложен и представлен вариант сервиса, объединяющий в себе алгоритмы для анализа данных.
- Описаны правила работы с построенным сервисом.
- Была обнаружена и исправлена ошибка в библиотеке PyOD (PR #108 – https://github.com/yzhao062/pyod/pull/108).
- Сервис доступен по адресу: http://bit.ly/anomd19.
- Код доступен по адресу: https://github.com/akarazeev/Bachelor.