

# Разработка механизма определения аномалий в данных

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

Научный руководитель:  
Дайняк А. Б.

# 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 уровня выброса.
- **Histogram-Based Outlier Score (HBOS)** – оценка выбросов на основе гистограммы.
- **Isolation Forest (IFOREST)** – метод изолирующего 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 используются 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.

# 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.

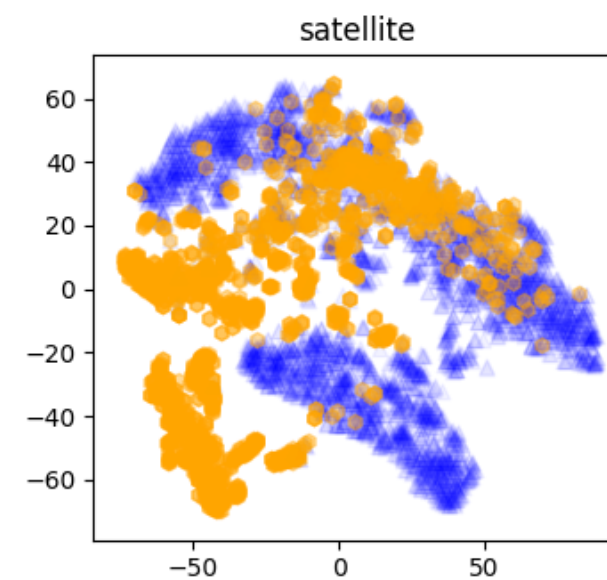
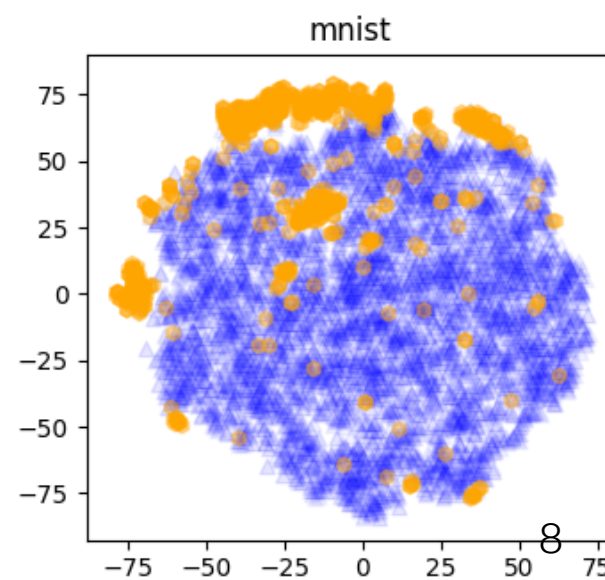
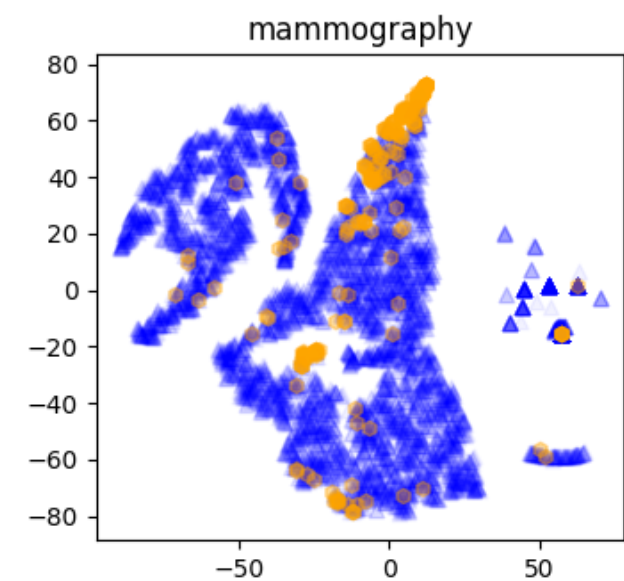
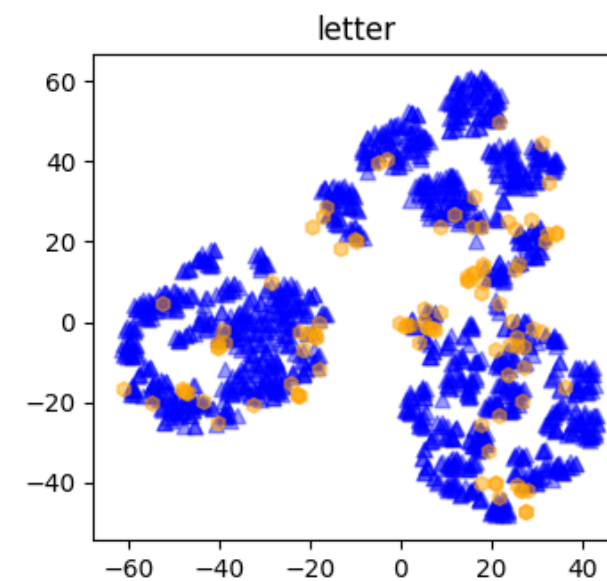
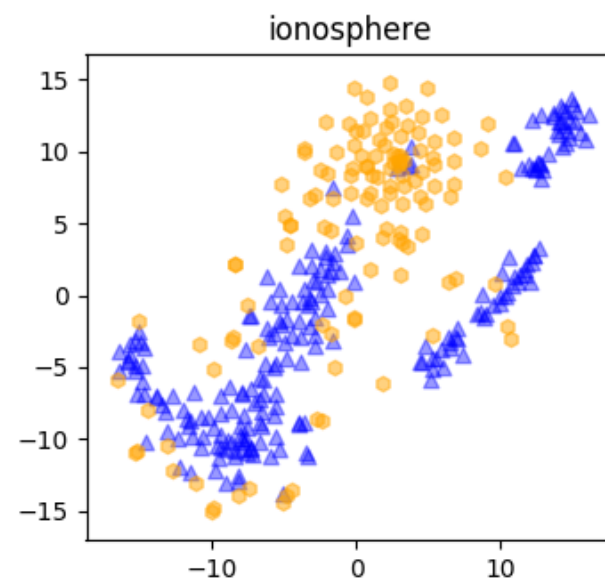
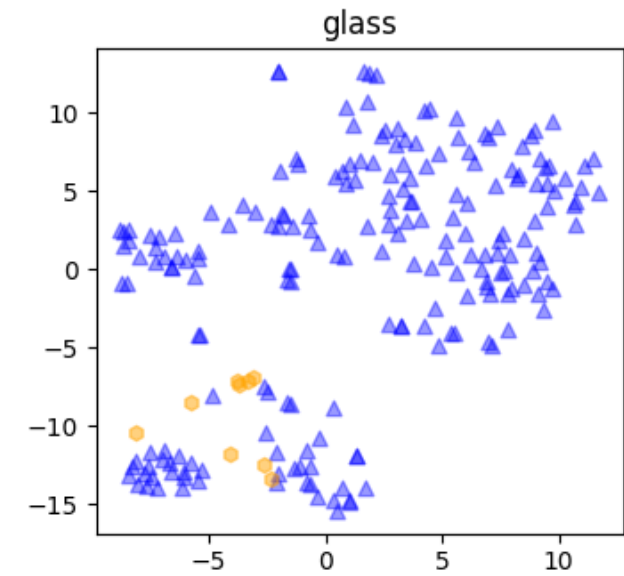
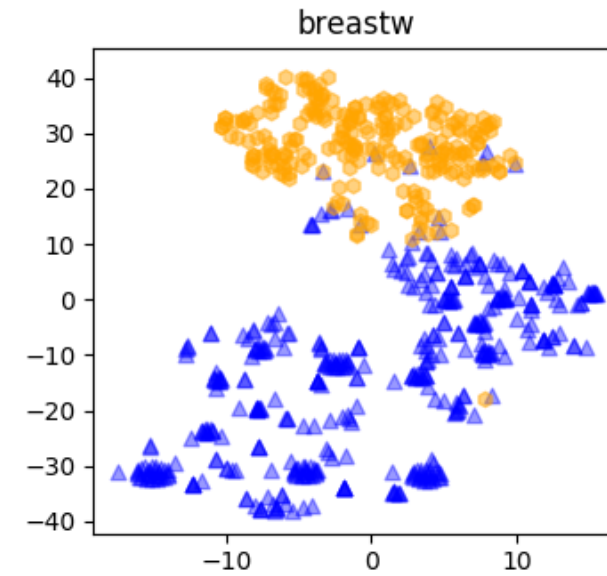
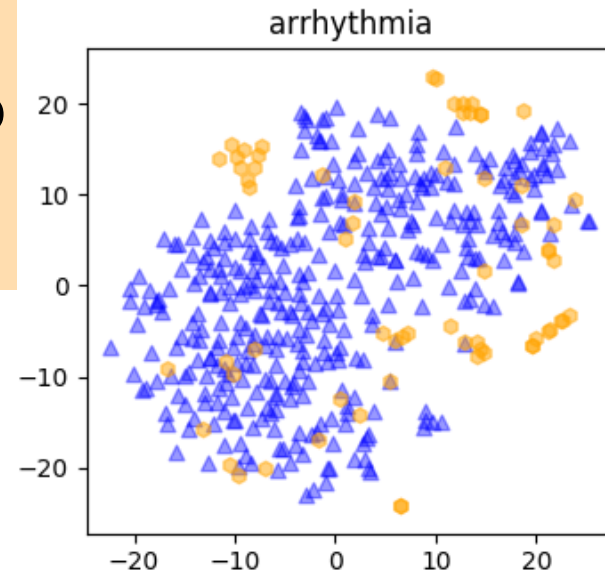
Датасет	KNN	PCA	OCSVM	LOF	HBOS	IFOREST
arrhythmia	0.7555	0.7794	0.7825	0.7672	0.7831	<b>0.7849</b>
breastw	<b>0.9908</b>	0.9608	0.9649	0.4574	0.9764	0.9872
glass	<b>0.8558</b>	0.7308	0.8077	0.6538	0.7500	0.7212
ionosphere	<b>0.9460</b>	0.8115	0.8684	0.9023	0.6190	0.8632
letter	<b>0.8660</b>	0.5119	0.5985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

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



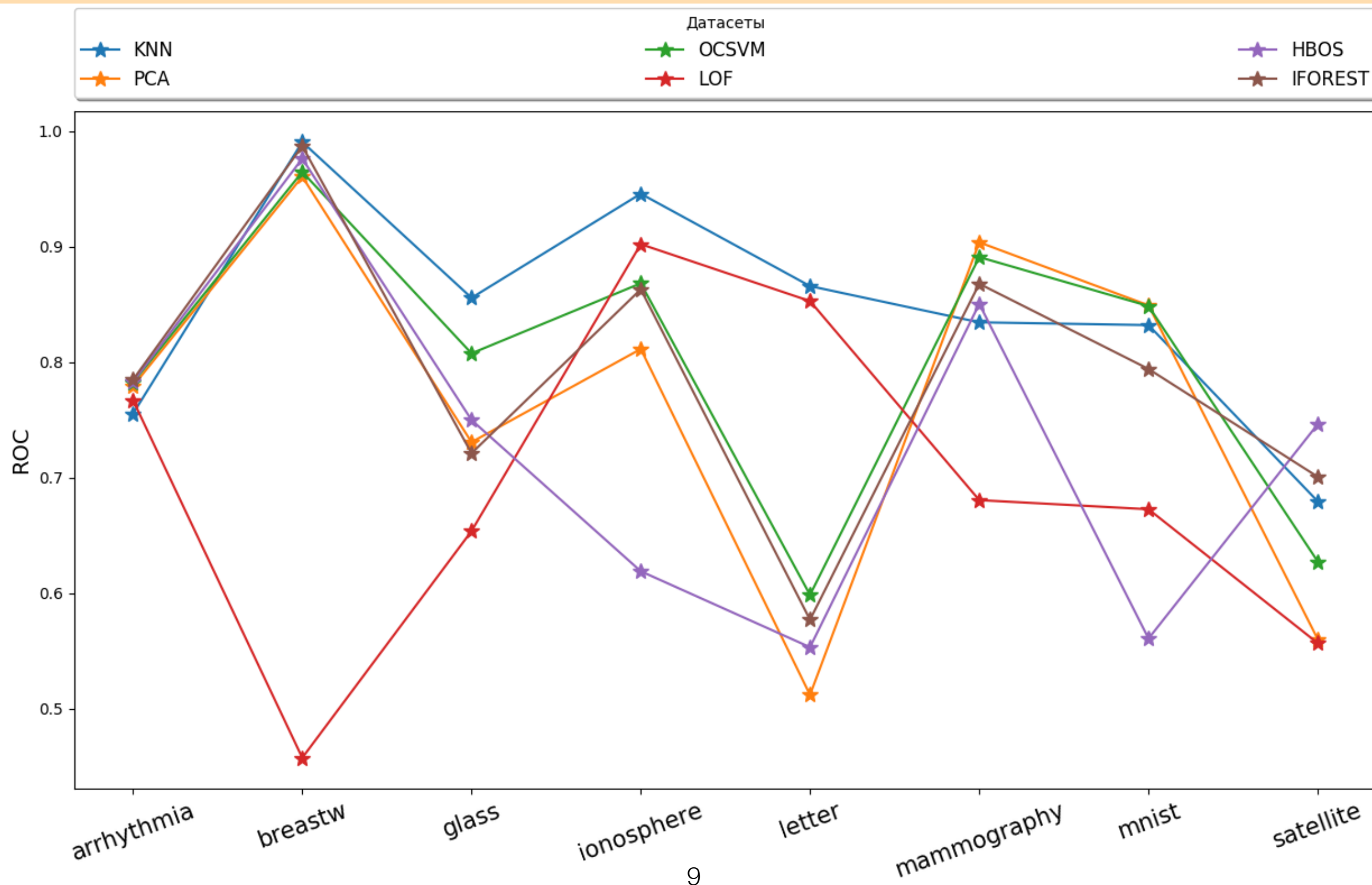
**Представление  
объектов из датасетов  
после понижения  
размерности с помощью  
алгоритма t-SNE**

Обозначения  
▲ Нормальные данные    ● Аномальные данные





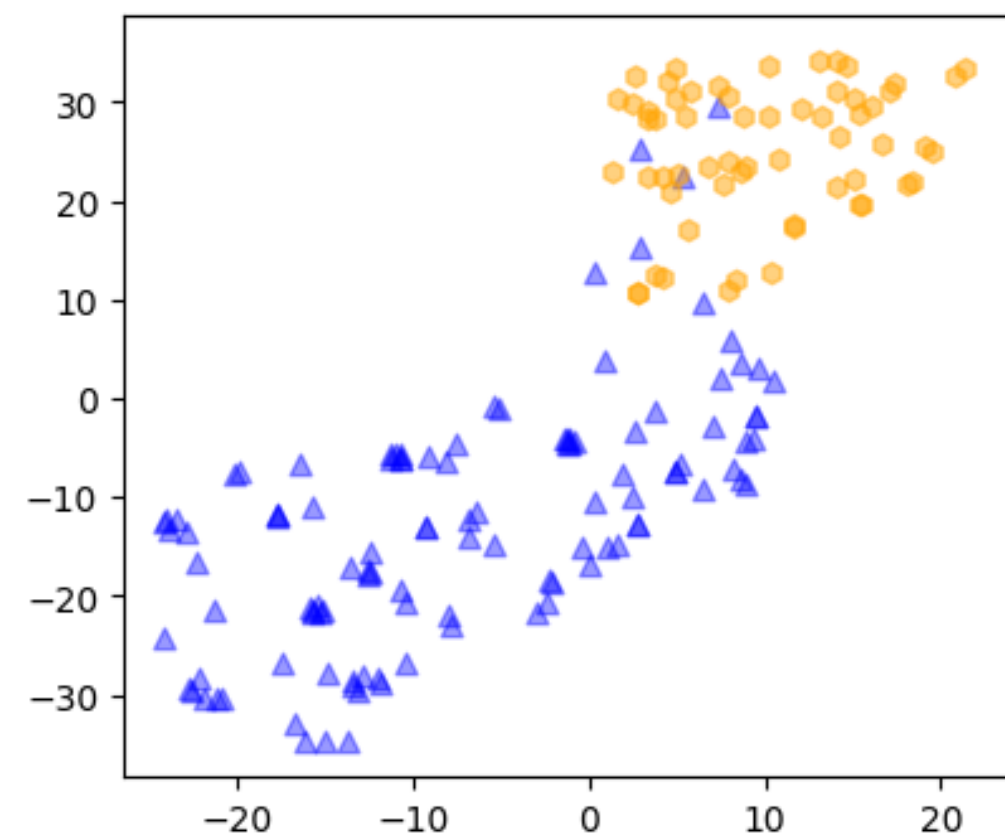
# Результаты обучения алгоритмов



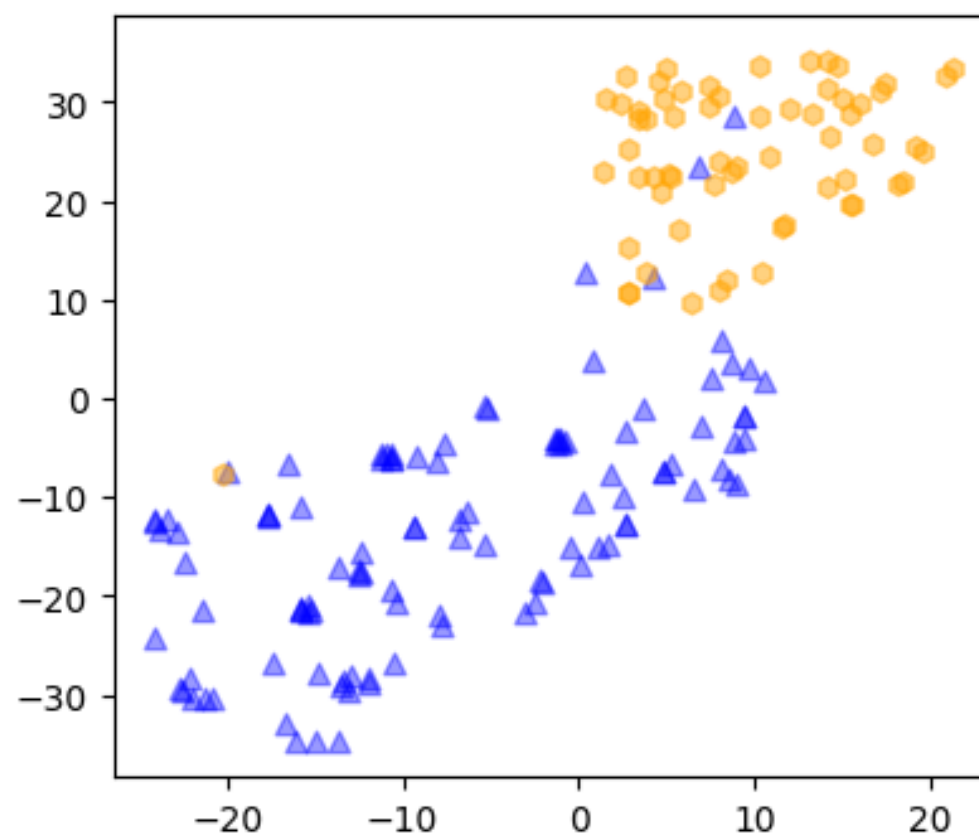
# Результаты обучения алгоритмов

Датасет: breastw, ROC: 0.9908  
Алгоритм: k-Nearest Neighbors

Ground truth



Predicted



Обозначения

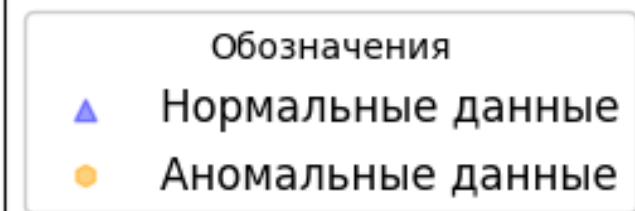
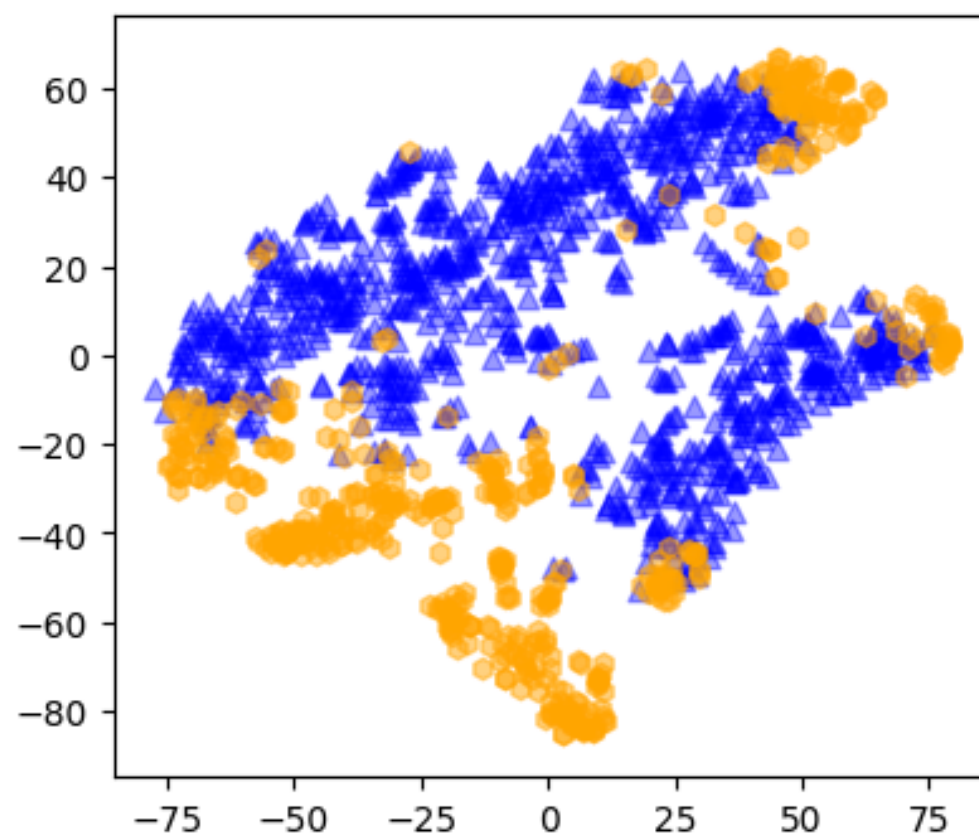
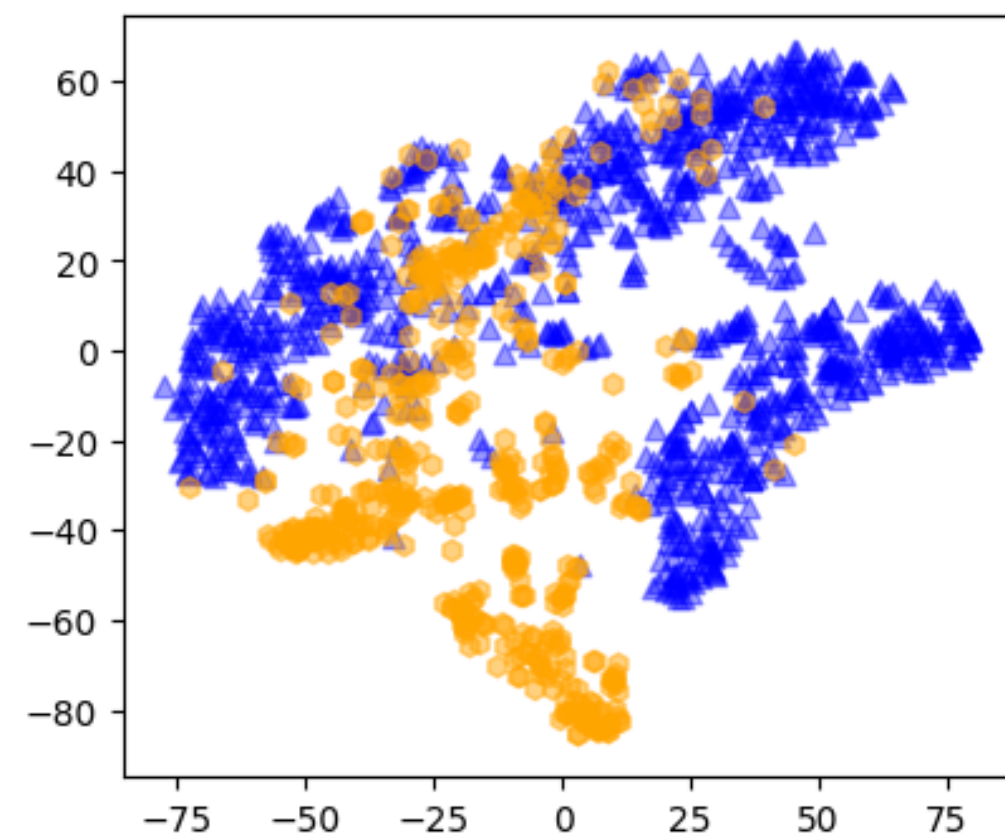
- ▲ Нормальные данные
- Аномальные данные

# Результаты обучения алгоритмов

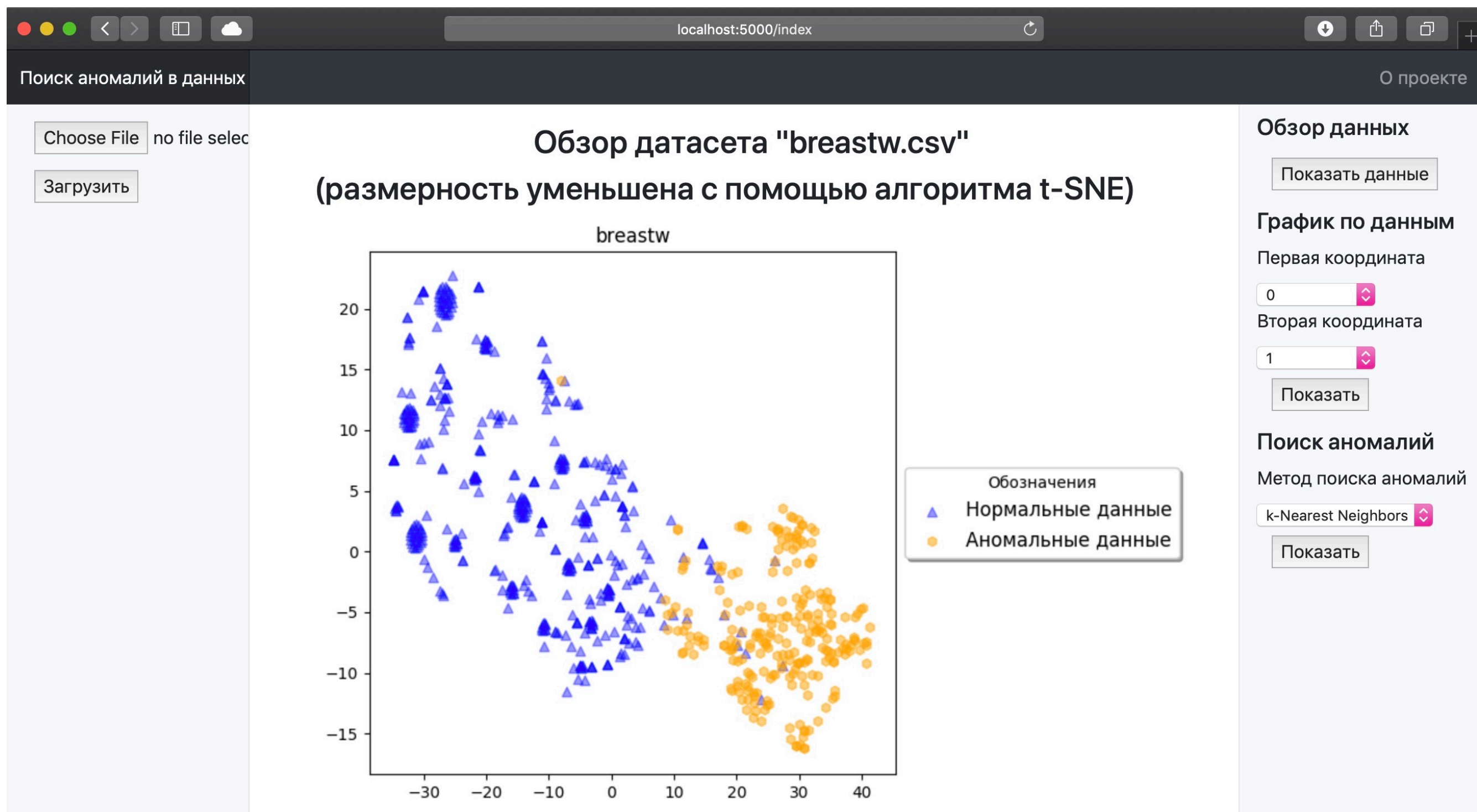
Датасет: satellite, ROC: 0.761  
Алгоритм: Histogram-Based Outlier Detection

Ground truth

Predicted

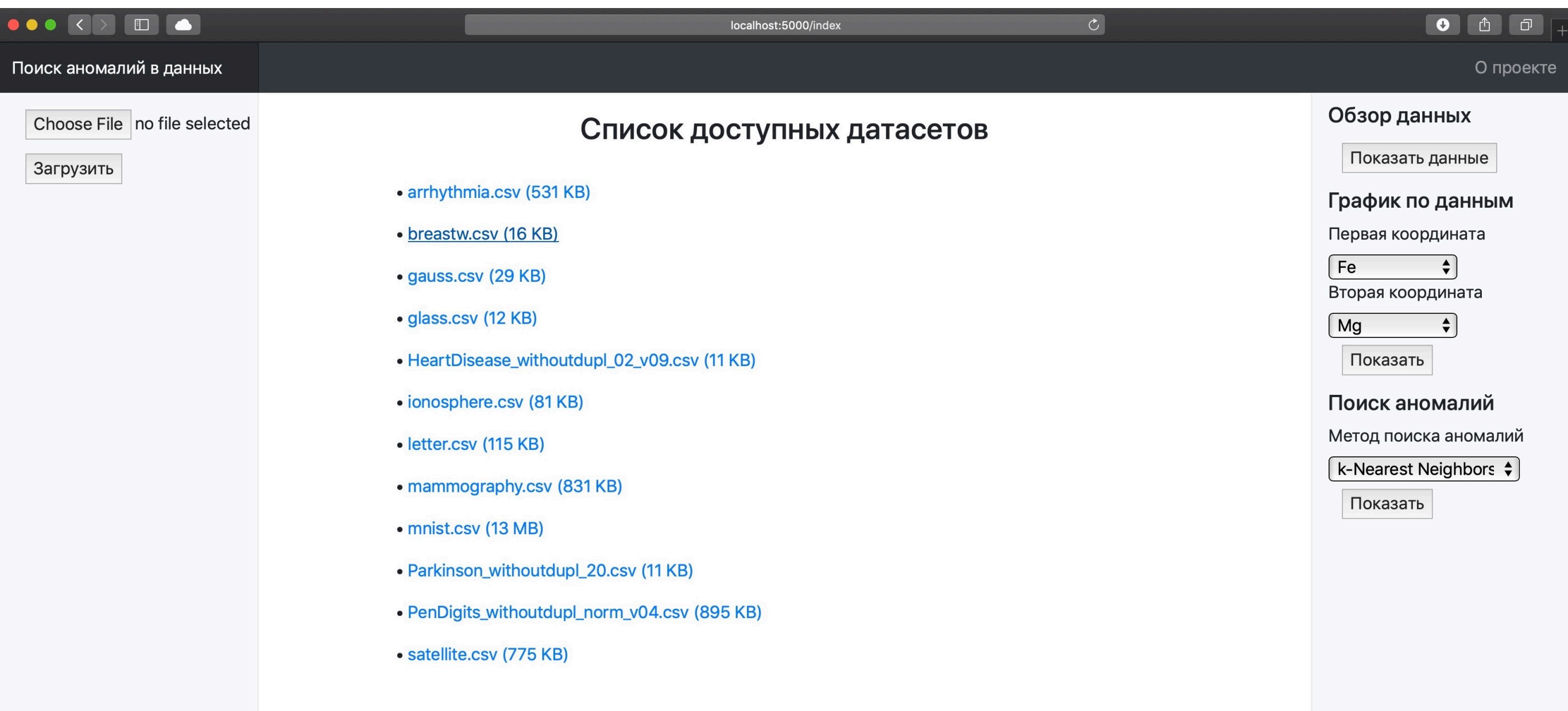


# Сервис



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

# Сервис



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



# Сервис

Поиск аномалий в данных

О проекте

Choose File

no file selected

Загрузить

Содержание датасета "glass.csv"

- Размерность данных: **9 + 1**
- Количество объектов: **214**
- Количество выбросов: **9 (4.21%)**

outlier	RI	Na	Mg	Al	Si	K
0.0	1.52101	13.64	4.49	1.1	71.78	0.06
0.0	1.51761	13.89	3.6	1.3599999999999999	72.73	0.48
0.0	1.51571	12.72	3.46	1.56	73.2	0.67
0.0	1.51763	12.8	3.66	1.27	73.01	0.6000000000000001
0.0	1.5158900000000002	12.88	3.43	1.4	73.28	0.6900000000000001
0.0	1.51748	12.86	3.56	1.27	73.21	0.54
0.0	1.51763	12.61	3.59	1.31	73.29	0.58
0.0	1.51618	13.53	3.55	1.54	72.99	0.39
0.0	1.51766	13.21	3.69	1.29	72.61	0.5700000000000001
0.0	1.51742	13.27	3.62	1.24	73.08	0.55
0.0	1.5159600000000002	12.79	3.61	1.62	72.97	0.64
0.0	1.51743	13.3	3.6	1.1400000000000001	73.09	0.58
0.0	1.51756	13.15	3.61	1.05	73.24	0.5700000000000001
0.0	1.51918	14.04	3.58	1.37	72.08	0.56
0.0	1.51755	13.0	3.6	1.3599999999999999	72.99	0.5700000000000001

Обзор данных

Показать данные

График по данным

Первая координата

RI

Вторая координата

Na

Показать

Поиск аномалий

Метод поиска аномалий

k-Nearest Neighbors

Показать

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>.