МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ФАКУЛЬТЕТ ИННОВАЦИЙ И ВЫСОКИХ ТЕХНОЛОГИЙ
КАФЕДРА КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Выпускная бакалаврская квалификационная работа на тему:

Разработка механизма определения аномалий в данных

Специальность 03.03.01 —
«Прикладные математика и физика (бакалавриат)»

Студент 4 курса, 592 гр.:
Каразеев Антон Андреевич

Научный руководитель:
Дайняк Александр Борисович

Москва — 2019

Оглавление

Стр.

Введение

В бакалаврской работе рассматриваются алгоритмы и методы определения аномалий в различных данных. Рассмотрена основная метрика оценки качества подобных алгоритмов – ROC-AUC.

Предлагается возможная реализация сервиса, который объединяет в себе современные методы по определению выбросов в данных.

Работа состоит из введения и четырёх глав. Полный объём работы составляет 27 страниц, включая 11 рисунков и 2 таблицы. Список литературы содержит 14 наименований.

Глава 1. Анализ предметной области

1.1 Постановка задачи

В настоящее время существует множество подходов к определению аномалий и выбросов в данных. Все они хорошо применимы в своей области. Именно поэтому следует в первую очередь проанализировать существующие алгоритмы и их сферы применимости. После чего предлагается объединить их в один сервис, чтобы упростить жизнь потенциальному пользователю, которому необходимо будет проанализировать большой объём данных.

1.2 Объект, предмет и методы исследования

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

1.3 Актуальность выбранной темы

Объём данных, которые появляются каждый день, растёт с экспоненциальной скоростью. Необходимо уметь работать с большими объёмами данных, чтобы получать из этого пользу. А для этого необходимо использовать актуальные алгоритмы и подходы, которые уже имеются.

Реализация современных алгоритмов поиска аномалий в данных может помочь в разных областях, таких как выявление отклонений в здоровье пациента, мошенничества в банковской сфере и др. Спрос на подобные сервисы в настоящее время высок и растёт с каждым днём.

Глава 2. Существующие алгоритмы

2.1 Анализ

В основном в качестве источника информации используются статьи [1—6]. Помимо статей интерес представляют датасеты, которые тоже предстояло найти. Среди датасетов есть хорошо известный MNIST с набором изображений рукописных цифр, а также данные о разных сортах винных изделий, свободных электронах в ионосфере Земли, о заболеваниях сердца и другие.

В этой работе рассматриваются стационарные данные. Поиск аномалий во временных рядах, а также прогнозирование временных рядов находятся за рамками рассматриваемой в работе темы.

2.2 Основы алгоритмов

Задачу поиска аномалий можно отнести к классу задач обучения без учителя. Суть поиска аномалий заключается в том, чтобы найти в выборке объекты, которые не похожи на большинство объектов выборки, т. е. те, которые выделяются на фоне других.

Часто бывает так, что аномальных объектов либо нет вообще, либо их очень мало и неизвестно где именно в выборке они находятся. Поэтому поиск аномалий относится к классу задач обучения без учителя (т. к. отсутствуют размеченные данные).

2.3 Результат работы алгоритмов

Результатом работы алгоритма для поиска аномалий могут быть как **степени аномалии** (anomaly scores), так и **бинарные метки** (binary labels).

В случае, когда алгоритм выдаёт **степень аномалии**, под степенью понимается уровень вероятности того, что объект является выбросом (аномалией).

В случае **бинарных меток** алгоритм сразу указывает на нормальные (обычно обозначаемые как 0) и аномальные (обозначаемые как 1) данные. Несмотря на то, что некоторые алгоритмы детектирования аномалий возвращают бинарные метки напрямую, степени аномалий тоже могут быть переведены в бинарное представление. 0 или 1 содержат меньше информации, чем степень аномалии. Тем не менее, это конечный результат, по которому обычно принимается решение об аномальности объекта выборки.

2.4 Классификация методов определения аномалий

Большинство методов определения аномалий используют метки, по которым можно определить, является ли объект выборки нормальным или аномальным. Поиск или сбор размеченных данных, которые будут точными и хорошо описывать рассматриваемую проблемы, чтобы хорошо обучить алгоритмы, довольно сложно и дорого.

Обычно выделяют три типа методов поиска аномалий:

1. **Supervised методы (обучение с учителем)**
   Предполагается, что имеется доступ к обучающим данным с точными и репрезентативными метками для нормальных и аномальных объектов. В таком случае обычно разрабатывают предсказательную модель для обоих классов. После обучения на тренировочных данных к каждому объекту из тестовой выборки применяется алгоритм, чтобы определить класс объекта.
2. **Semi-supervised методы (обучение с частичным привлечение учителя)**
   Предполагается, что имеются размеченные данные только для нормального класса. Так как для обучения таких алгоритмов не требуются размеченные аномальные данные, они имеют более широкое применение, чем supervised методы.
3. **Unsupervised методы (обучение без учителя)**

Такие методы не требуют обучающих данных и поэтому наиболее широко используются. Unsupervised методы поиска аномалий могут нормальные данные из всех представленных и рассматривать отклонение от них как аномалию.

Многие semi-supervised методы могут быть использованы для unsupervised случая. Например, с их помощью можно дополнительно семплировать объекты из выборки, если данных для обучения алгоритма попросту недостаточно.

2.4.1 Метрика качества модели

В качестве основной метрики используется ROC-AUC, что расшифровывается как receiver operating characteristic – area under curve и переводится как ”рабочая характеристика приёмника – площадь под кривой”. ROC-кривая позволяет оценить качество бинарной классификации[1]. Она показывает соотношение между долей объектов, которые были верно классифицированы как принадлежащие определённому классу (True Positive Rate, TPR), от общего числа объектов в выборке Именно площадь под кривой и выступает в роли метрики качества. Минимальное значение, которое может принимать ROC-AUC составляет 0.5, а максимальное – 1.

График на рисунке 2.1 был построен с помощью библиотеки scikit-learn[2].

[1] бинарная означает, что есть лишь два класса объектов – например, нормальные и аномальные

[2] https://scikit-learn.org/0.15/auto_examples/plot_roc.html

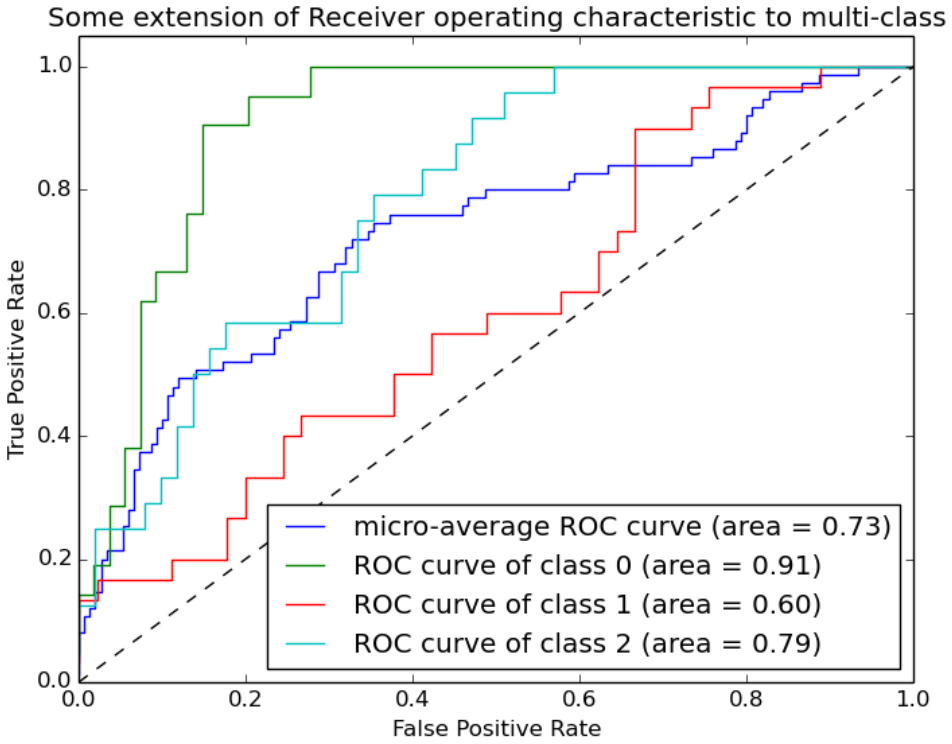

Рисунок 2.1 — Пример ROC-кривой.

2.5 Подходы к решению

Одним из возможных способов определения аномалий является измерение схожести между объектов. У такого способа есть два варианта:

1. Восстановление плотности
2. Классификация

2.5.1 Восстановление плотности

В случае с восстановлением плотности необходимо построить распределение, которое хорошо описывает выборку. И это распределение позволяет посчитать вероятность для нового объекта получить его из распределения, описывающего выборку.

В терминах этого метода аномалия - объект, полученный из другого распределения, описывающего другую выборку данных.

Есть три подхода:

1. Параметрический
2. Непараметрический
3. Восстановление смесей

Параметрический метод – Распределение представляется в виде $p(x) = \varphi(x|\theta)$, где θ выступает в качестве параметра распределения. Например, в семейство параметрических распределений входит распределение Гаусса - $\theta = (\mu, \Sigma)$, где μ - вектор средних и Σ - ковариационная матрица.

Параметры модели подбираются таким образом, чтобы вероятность объектов из обучающей выборки была максимальной. Для этого обычно пользуются Методом Максимального Правдоподобия:

$$\sum_i \log \varphi(x_i|\theta) \to \max_\theta$$

2.6 Алгоритмы

В текущей работе были рассмотрены следующие алгоритмы:

1. k-Nearest Neighbors (k-NN) [7] – метод k-ближайших соседей.
2. Principal Component Analysis (PCA) [8] – метод главных компонент.
3. One-Class Support Vector Machines (OCSVM) [9] – одноклассовый метод опорных векторов.
4. Local Outlier Factor (LOF) [10] – метод локального уровеня выброса.
5. Histogram-Based Outlier Score (HBOS) [11] – оценка выбросов на основе гистограммы.
6. Isolation Forest [12] – метод изолируещего леса.

2.7 Датасеты

Для проверки сервиса были рассмотрены следующие датасеты:

1. Arrhythmia – определение наличия аритмии по данным ЭКГ [13].
2. Breast Cancer – определение типа опухоли молочной железы: доброкачественная или злокачественная.
3. Glass – идентификация типа стекла, оставленного на месте преступления.
4. Ionosphere – рассматриваются характеристики радаров, которые используется в анализе ионосферы: необходимо определить является радар ”плохим”или ”хорошим”.
5. Letter Recognition – по описанию изображения определить присутствует ли буква из английского алфавита или нет.
6. Mammography – детектирование микрокальцинатов по данным маммографии.
7. MNIST – научиться различать изображения рукописных цифр 6 и 0.
8. Satellite – определение типа почвы по спутниковым снимкам.

В качестве источника этих данных выступает библиотека ODDS [14].

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

Сравнение данных, которые представлены в указанных датасетах.

При построении графиков, которые изображены на рисунке 2.3, использовалась библиотека adjustText[3].

[3] https://github.com/Phlya/adjustText/

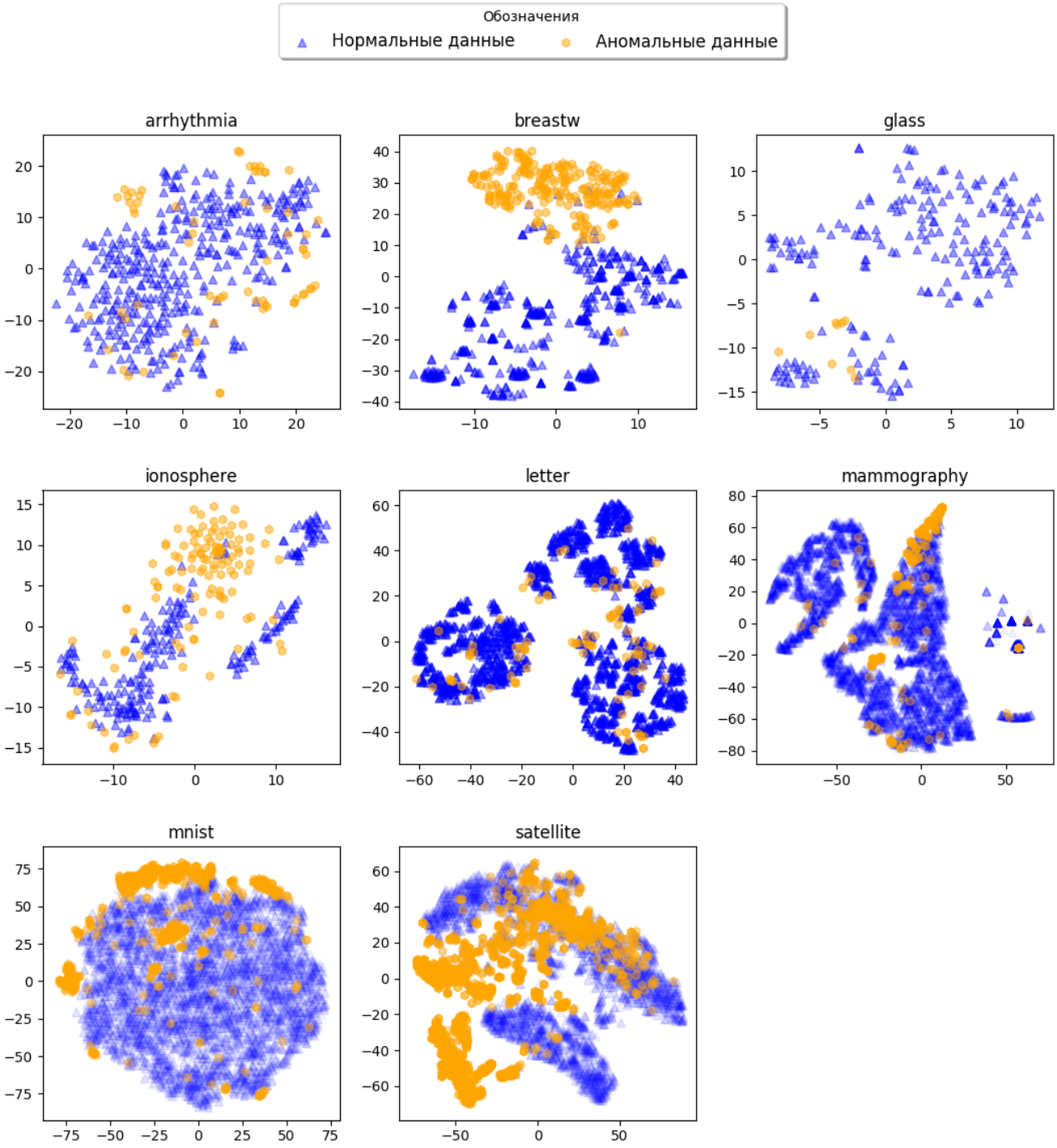

Рисунок 2.2 — Рассматриваемые датасеты после применения алгоритма понижения размерности t-SNE.

Демонстрация наилучших результатов работы алгоритмов (рисунок 2.5).

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

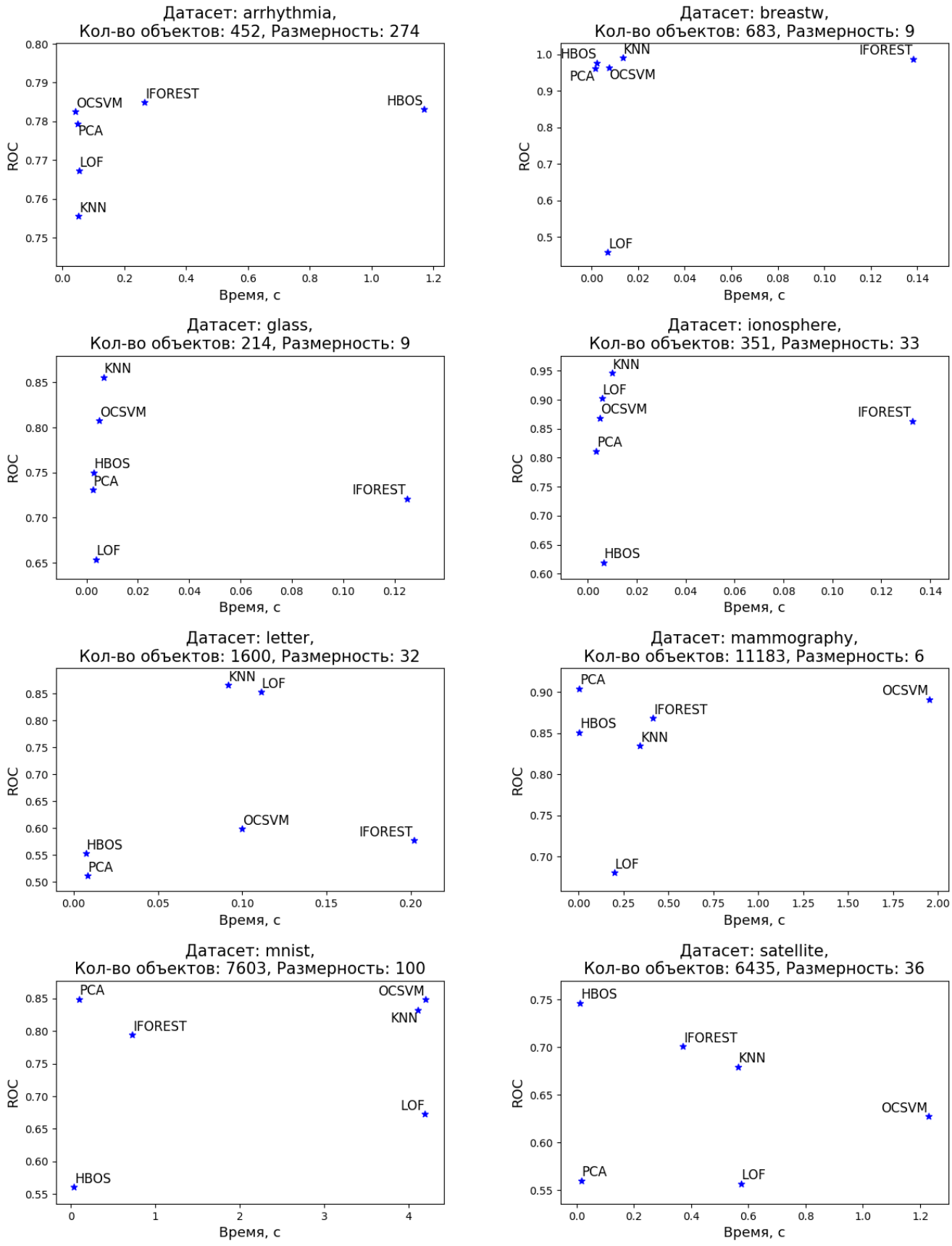

Рисунок 2.3 — Эффективность алгоритмов на разных датасетах.

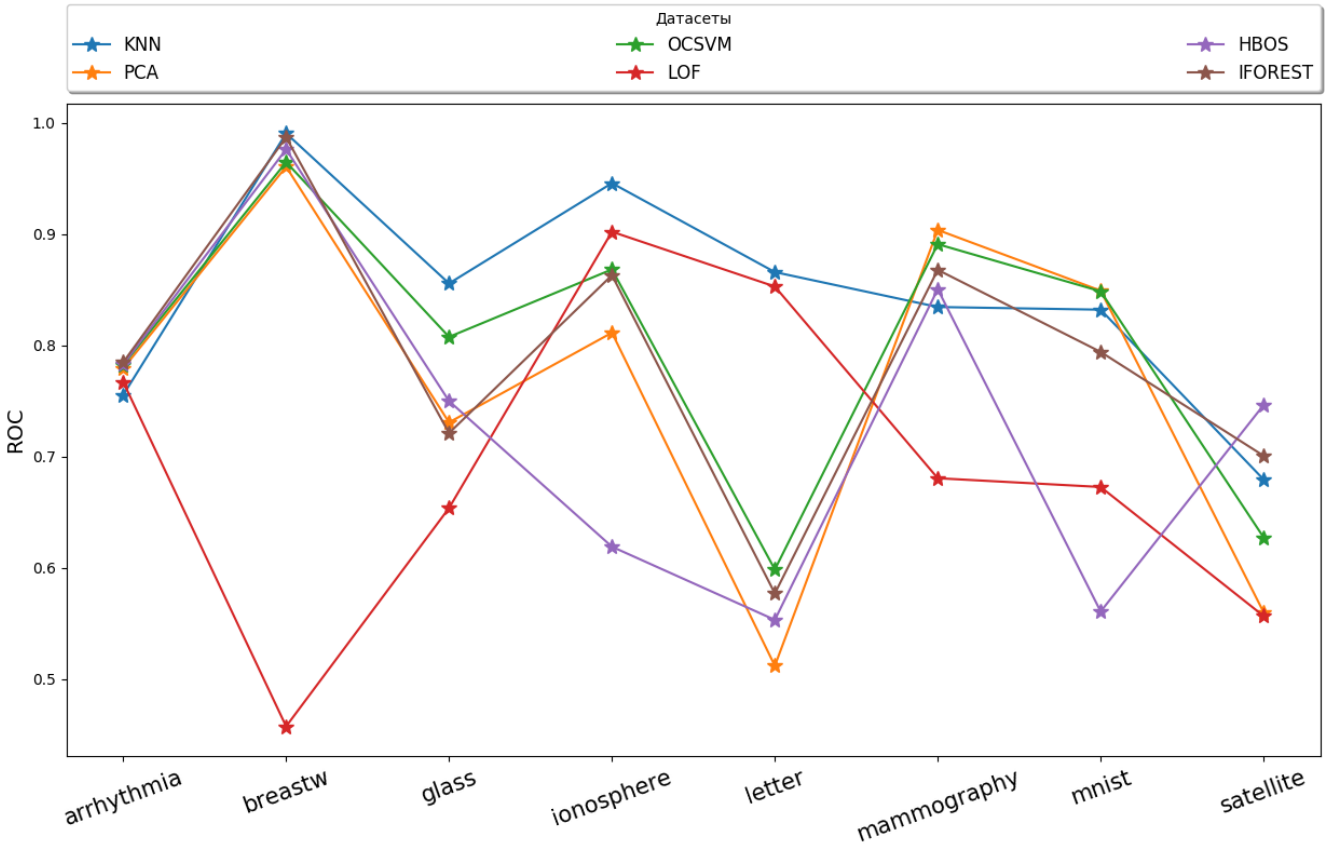

Рисунок 2.4 — Качество алгоритмов в зависимости от датасета.

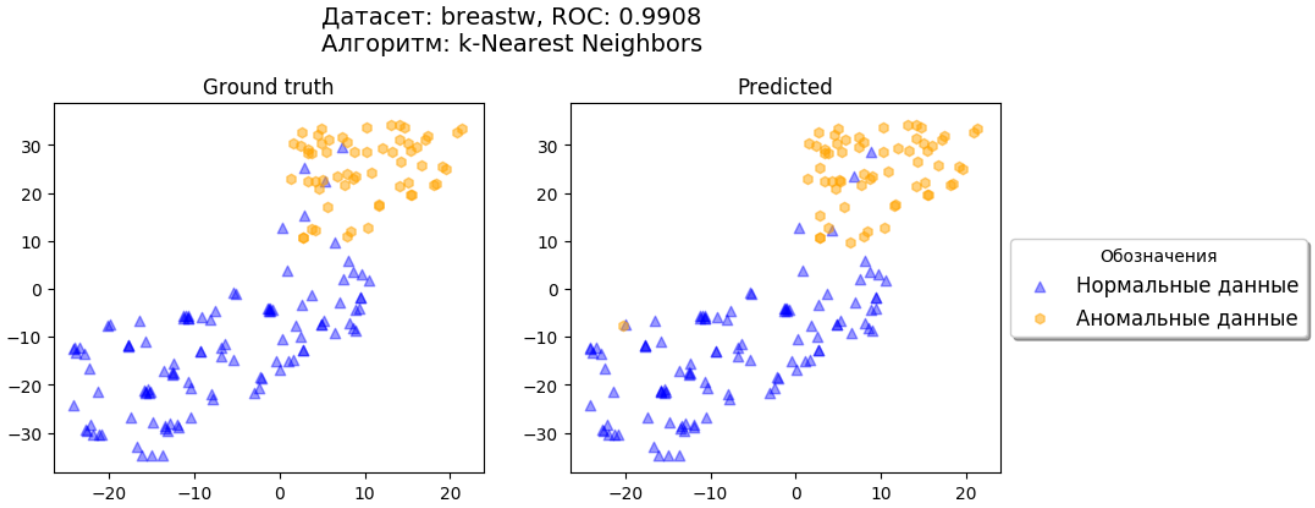

Рисунок 2.5 — Датасет Breast Cancer.

Глава 3. Сервис для анализа данных

В этой главе будет рассмотрено устройство сервиса, которое было создано, чтобы упростить процесс анализа данных и поиска аномалий в них.

Ниже будут представлены основные сценарии работы с сервисом.

3.1 Описание сервиса

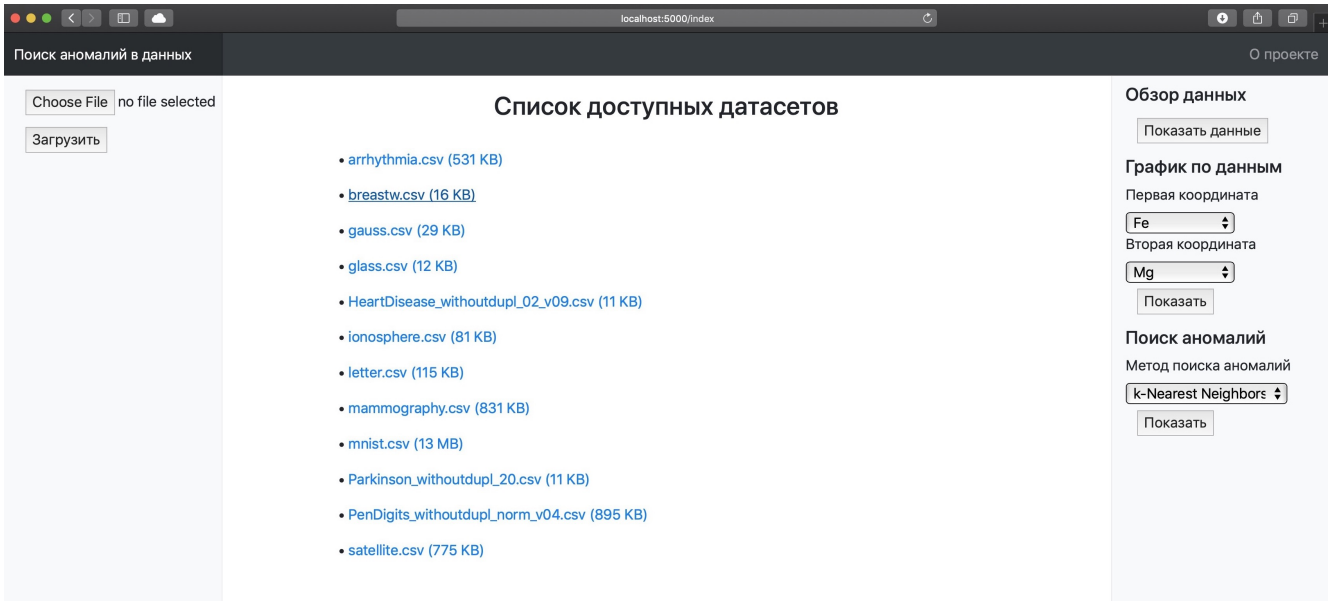

Рисунок 3.1 — Главный экран сервиса. Есть возможность загрузить данные через окно загрузки (как продемонстрировано рисунке 3.2), либо выбрать датасет из списка загруженных ранее.

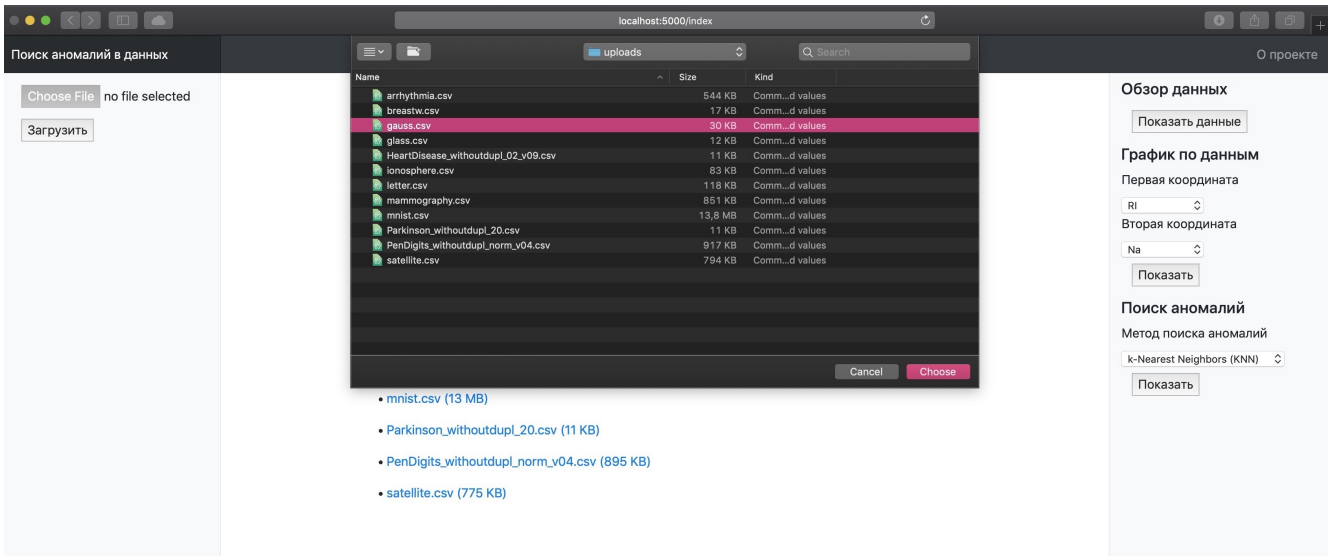

Рисунок 3.2 — Пример экрана загрузки датасета на сервер.

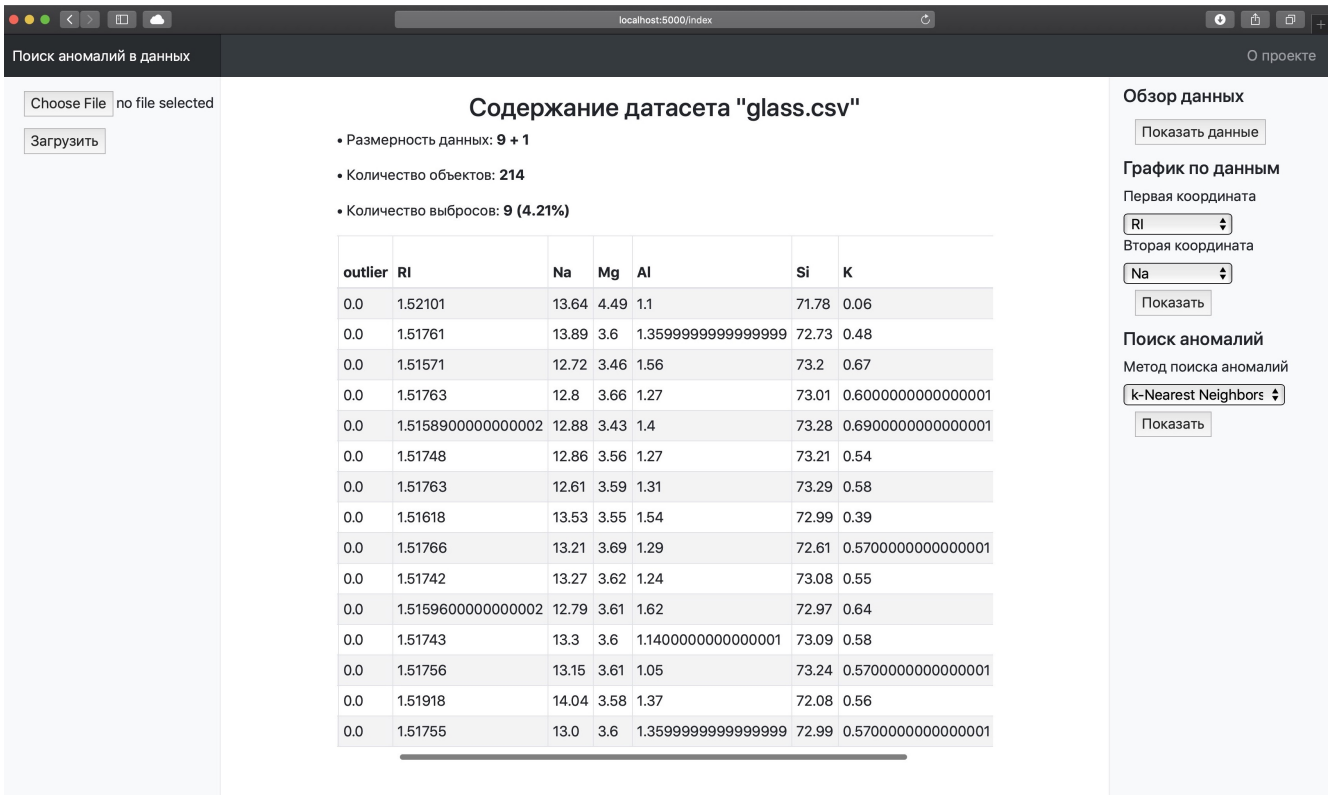

Рисунок 3.3 — После выбора файла на экране появляется таблица с некоторыми объектами из указанного датасета. С помощью меню, которое располагается справа, можно выбрать: **(i)** построение общего графика по датасету, на котором будут представлены данные, размерность которых была понижена с помощью метода t-SNE (продемонстрировано на рисунке 3.4), **(ii)** график зависимости одной из размерностей данных от другой (продемонстрировано на рисунке 3.5) и **(iii)** выбрать алгоритм из списка, с помощью которого будет осуществлён поиск выбросов в данных (продемонстрировано на рисунке 3.6).

3.2 Основные функции

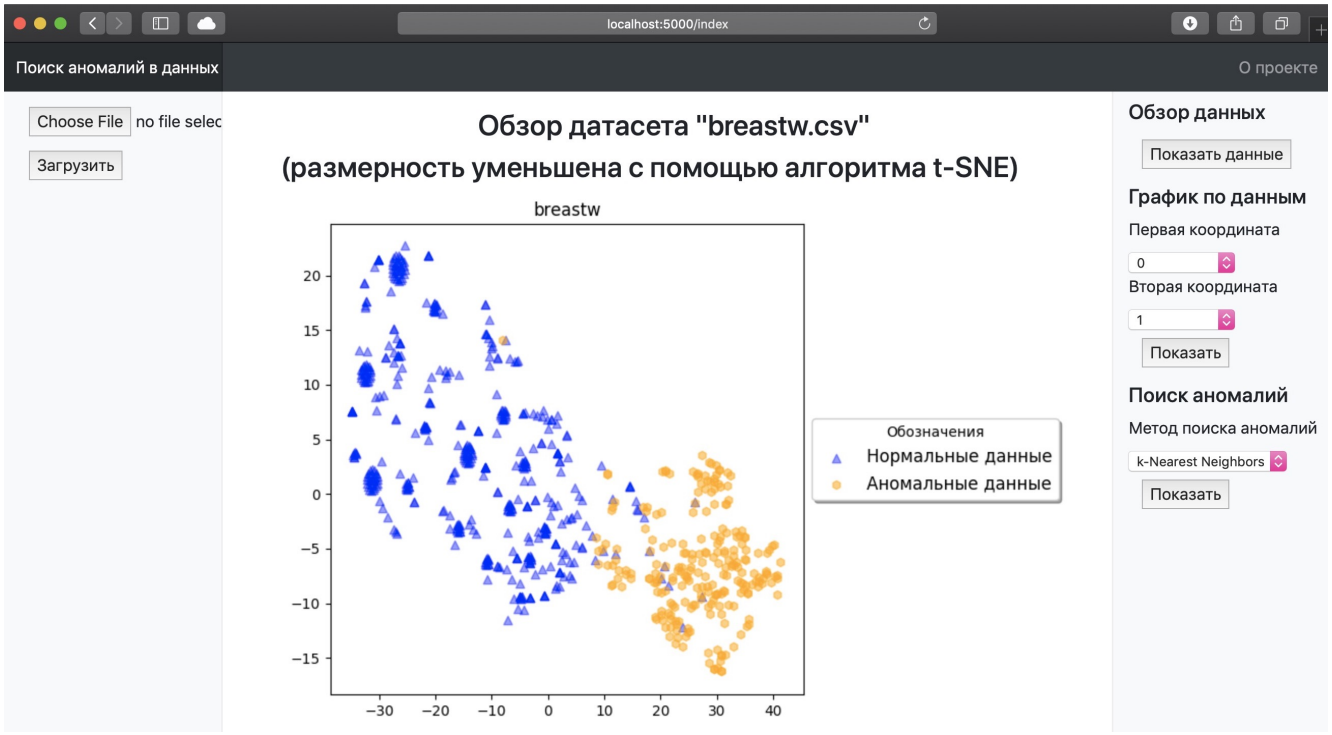

Рисунок 3.4 — Представление данных из датасета после понижения размерности до dim=2 при помощи алгоритма t-SNE.

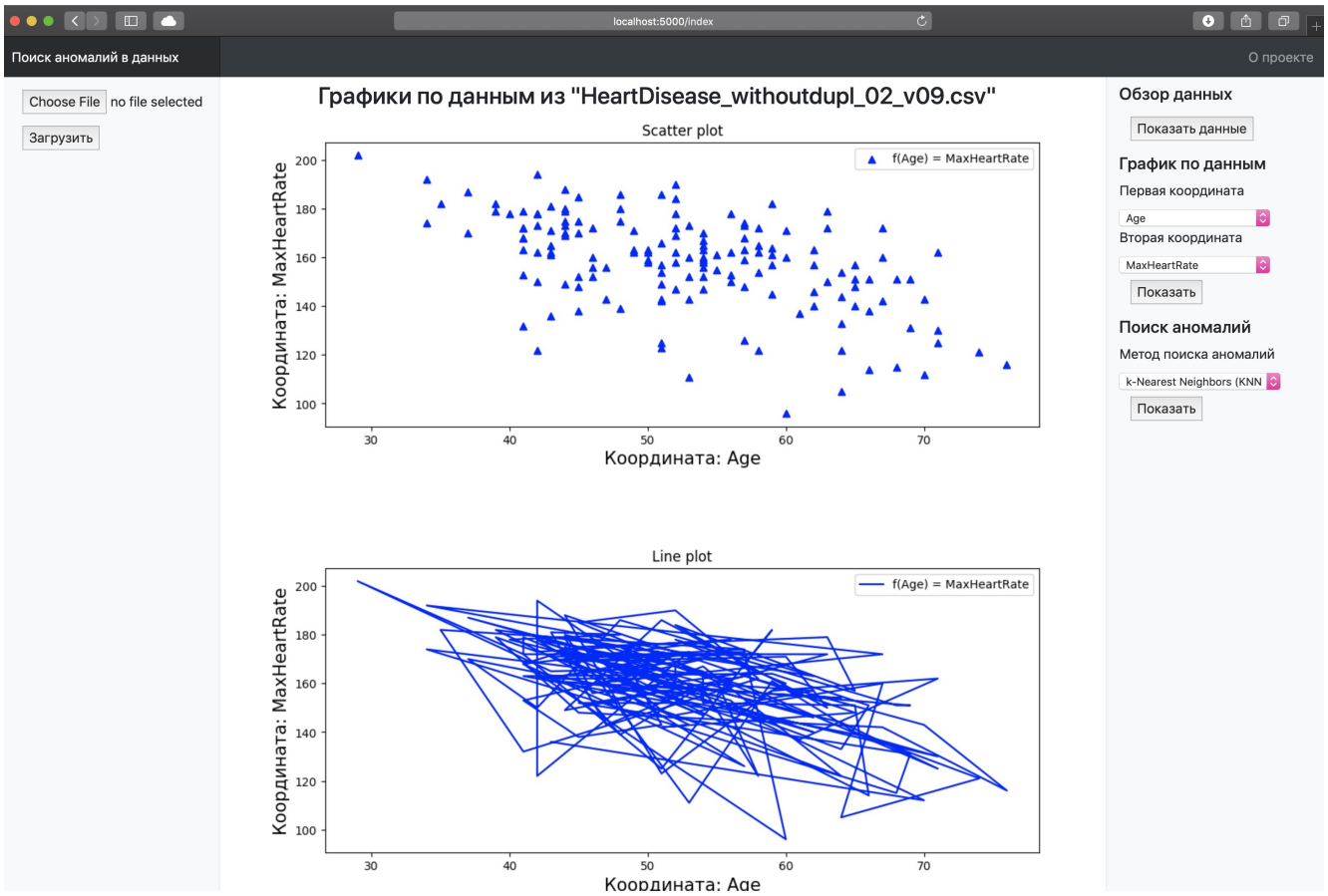

Рисунок 3.5 — Построение графика зависимости для указанных координат. На рисунке представлена зависимость максимального количества ударов в минуту от возраста пациентов (данные из датасета сердечных заболеваний).

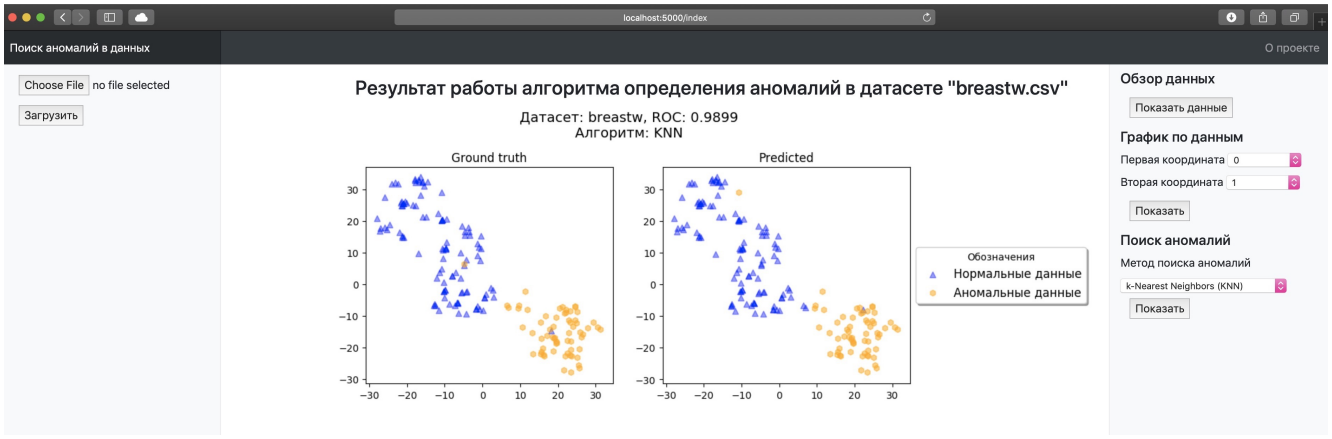

Рисунок 3.6 — Поиск выбросов в данных. Датасет разбивается на две непересекающиеся части: на первой части выбранная модель обучается, а на второй – проверяется качество обученной модели. Обычно, разделение происходит в соотношении 75% и 25% соответственно. Так снижается вероятность переобучения на данных, что позволяет избежать ухудшения обобщающей способности алгоритма.

Глава 4. Результаты

4.1 Выводы

В ходе работы были проанализированы существующие библиотеки для языка программирования Python, которые упрощают работу с алгоритмами для поиска аномалий и выбросов в данных.

Была обнаружена и исправлена ошибка в библиотеке PyOD[1].

Построен сервис для анализа данных и определения аномалий. Сервис развёрнут на удалённом сервере с операционной системой Ubuntu 18.04.2 LTS (Bionic Beaver)[2].

В сервисе используется стандартный подход сервер-клиент. В качестве сервера выступает база данных PostgreSQL[3]. С помощью библиотеки Flask[4] для языка программирования Python было построено веб-приложение – на данный момент именно оно выступает в роли клиентского приложения, через который можно получить доступ к функционалу всего сервиса.

Помимо прочего, понадобилось дополнительное время, чтобы развернуть всю систему на сервере и добиться корректной работы. На момент написания этой квалификационной работы сервис был запущен по адресу http://d6719ff8.ngrok.io/index.

[1] Pull Request #108 в репозитории PyOD – https://github.com/yzhao062/pyod/pull/108
[2] http://releases.ubuntu.com/18.04/
[3] https://www.postgresql.org/docs/11/
[4] http://flask.pocoo.org

Список литературы

1. *Dai*, *W.* Directional Outlyingness for Multivariate Functional Data / W. Dai, M. G. Genton. — 2018. — URL: https://stsda.kaust.edu.sa/Documents/2019.DG.CSDA.pdf.
2. *Hodge*, *V. J.* A Survey of Outlier Detection Methodologies / V. J. Hodge, J. Austin. — 2004. — URL: https://www-users.cs.york.ac.uk/vicky/myPapers/Hodge+Austin_OutlierDetection_AIRE381.pdf.
3. *Vakili*, *K.* Finding Multivariate Outliers With FastPCS / K. Vakili, E. Schmitt. — 2013. — URL: https://arxiv.org/pdf/1301.2053.pdf.
4. *Chandola*, *V.* Anomaly Detection: A Survey / V. Chandola, A. Banerjee, V. Kumar. — 2009. — URL: https://www.vs.inf.ethz.ch/edu/HS2011/CPS/papers/chandola09_anomaly-detection-survey.pdf.
5. *Billor*, *N.* BACON: blocked adaptive computationally efficient outlier nominators / N. Billor, A. S. Hadi, P. F. Velleman. — 2000. — URL: https://www.sciencedirect.com/science/article/pii/S0167947399001012.
6. *Wilkinson*, *L.* Visualizing Big Data Outliers through Distributed Aggregation / L. Wilkinson. — 2017. — URL: https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf.
7. *Ramaswamy*, *S.* Efficient Algorithms for Mining Outliers from Large Data Sets / S. Ramaswamy, R. Rastogi, K. Shim. — 2000. — URL: https://dl.acm.org/citation.cfm?id=335437.
8. A Novel Anomaly Detection Scheme Based on Principal Component Classifier / M.-L. Shyu [et al.]. — 2003. — URL: https://www.researchgate.net/publication/228709094_A_Novel_Anomaly_Detection_Scheme_Based_on_Principal_Component_Classifier.
9. Estimating the support of a high-dimensional distribution / B. Schölkopf [et al.]. — 2001. — URL: https://eprints.soton.ac.uk/259007/1/TRONECLA.PS.
10. LOF: Identifying Density-Based Local Outliers / M. M. Breunig [et al.]. — 2000. — URL: http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf.

11. *Goldstein*, *M.* Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm / M. Goldstein, A. Dengel. — 2012. — URL: https://pdfs.semanticscholar.org/5cf8/81d1db19834f123fcfc79ad32097aeafe17f.pdf.

12. *Liu*, *F. T.* Isolation Forest / F. T. Liu, K. M. Ting, Z.-H. Zhou. — 2008. — URL: https://ieeexplore.ieee.org/abstract/document/4781136.

13. A Supervised Machine Learning Algorithm for Arrhythmia Analysis / H. A. Guvenir [et al.]. — 1997. — URL: http://repository.bilkent.edu.tr/bitstream/handle/11693/27699/bilkent-research-paper.pdf?sequence=1.

14. *Rayana*, *S.* ODDS Library / S. Rayana. — 2016. — URL: http://odds.cs.stonybrook.edu.

Список рисунков

Список таблиц