Д. А. Фоменко[1], В. Д. Шулепова[1],
В. В. Ожирельев[2] , А. А. Ступников[1]
[1]Тюменский государственный университет, г. Тюмень
[2]Тюменский государственный медицинский университет, г. Тюмень
УДК 614.1:311

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ РЕЗУЛЬТАТОВ МЕДИЦИНСКИХ ЛАБОРАТОРНЫХ АНАЛИЗОВ

Аннотация. Данная работа рассматривает проблему хранения и обработки результатов медицинских лабораторных анализов. Медицинские учреждения хранят данные в базах данных с отличающимися структурами, и отсутствует специализированное программное обеспечение для статистического анализа накопленных данных. В качестве решения создано веб-приложение, предлагающее инструменты статистического анализа с их визуализацией и способное агрегировать выгрузки из медицинских учреждений.

Ключевые слова: веб-разработка, статистический анализ, кластерный анализ, обработка медицинских данных.

Введение. В Тюменской области последние несколько лет идет активная цифровизация в сфере здравоохранения, которая позволила накопить большой объем информации о пациентах и их лабораторных анализах. Собранные данные могли бы использоваться медицинскими исследователями, однако это затруднительно для них из-за ряда проблем.

На текущий момент информация о результатах лабораторных анализов не хранится в единой базе данных — данные хранятся в каждом районе Тюменской области в своей собственной базе данных, структуры и справочники которых не совпадают между собой. Нет информации о том, чтобы департамент здравоохранения Тюменской области анонсировал проекты по созданию единой базы медицинских данных. Также не существует специализированного ПО для обработки данных о лабораторных анализах.

Исходя из этого, актуальна разработка программного обеспечения, представляющее собой единую систему агрегации и обработки результатов медицинских лабораторных анализов.

Целью работы является создание специализированного приложения для агрегации и статистической обработки данных о результатах лабораторных анализов и разработка базы данных, в которую могут быть импортированы данные с разным форматом их представления.

Множество статистических методов, которые часто используются в медицинских исследованиях, и их описание представлено в книгах [1, 2]. Данные ресурсы наполнены основами проведения статистического анализа на медицинских данных — эти материалы дают нам общее представление о том, какие методы стоит рассматривать для возможной последующей реализации в разрабатываемом приложении.

В монографии Меймана и Рокоча [3] представлен обзор основных используемых для выявления закономерностей методов кластеризации и математических методов, лежащих в их основе. Подробно рассмотрены метрики расстояния для параметров разных типов. Также затрагиваются проблемы выполнения кластеризации на больших наборах данных и определения количества кластеров.

Основы методов оценки значимости статистических гипотез и их реализации на Python рассмотрены в пособии [4].

Для поиска и удаления выбросов из выборок часто используется метод local outlier factor (LOF), предложенный и описанный в статье [5].

Так как визуализация результатов методов статистического анализа важна для разрабатываемого приложения, были рассмотрены материалы с описанием корректного использования способов визуализации. В медицине есть правила выбора вида графиков или диаграмм, о которых ведется речь в статье [6], где приведена схема выбора способа графического представления результатов в зависимости от типа анализируемых данных и цели анализа данных.

Материалы и методы. Исходные наборы данных с результатами лабораторных анализов, проведенных в Ишимском и Кондинском районах Тюменской области, были получены в рамках исследования, проводимого в Тюменском государственном медицинском университете.

Набор данных из Ишимского района представляет собой один файл формата csv, содержащий 1 028 766 строк. Набор данных из Кондинского района представляет собой 13 файлов в формате csv, суммарно содержащие 658 462 строк.

Данные файлы являются выгрузками из системы 1С:Медицина и различаются как по количеству колонок, так и по их содержимому:

1. Значения в сходных колонках часто представлены в разных форматах.

2. Названия тестов, анализов и их идентификаторы в датасетах не совпадают (некоторые названия встречаются в виде аббревиатуры в одной выгрузке и в виде полного названия в другой; иногда наименования приведены на английском языке), а также имеют орфографические ошибки.

3. В датасете Кондинского района частично отсутствует идентификатор направления.

Для реализации задач медико-биологического исследования данных медицинских анализов в разработанную систему были включены модели и методы статистической обработки и исследования данных:

1) иерархическая кластеризация, кластеризация k-средних, метод локального уровня выброса, метод главных компонент для выявления закономерностей;

2) t-критерии, медианный критерий, критерий χ-квадрат и дисперсионный анализ для оценки статистических гипотез.

Так как в структуре БД присутствуют реляционные связи между сущностями для организации базы данных необходима реляционная СУБД, в качестве которой была выбрана PostgreSQL.

Для серверной части приложения был использован Python в сочетании с фреймворком Flask, который предоставляет функционал для работы сервера. Были использованы статистические библиотеки numpy, pandas, scipy, scikit-learn, pyclustering. Работа с базой данных осуществляется с помощью библиотеки psycopg2.

Для клиентской части проекта были использованы следующие технологии: HTML, CSS, JavaScript, React (JavaScript фреймворк для управления состоянием клиентской части приложения).

Результаты

Архитектура веб-приложения

Архитектура приложения изображена на рис. 1. При первоначальной загрузке страницы пользователю будут отправлены данные о доступных анализах и диагнозах. Затем пользователь должен сформировать выборки и указать интересующий его инструмент — эта информация попадает в соответствующую часть модуля проведения исследований.

Модуль получения данных используются для сокрытия деталей работы с БД. Он принимает фильтры выборок и отсылает саму выборку в формате таблицы DataFrame. Данные в БД попадают из модуля импорта.

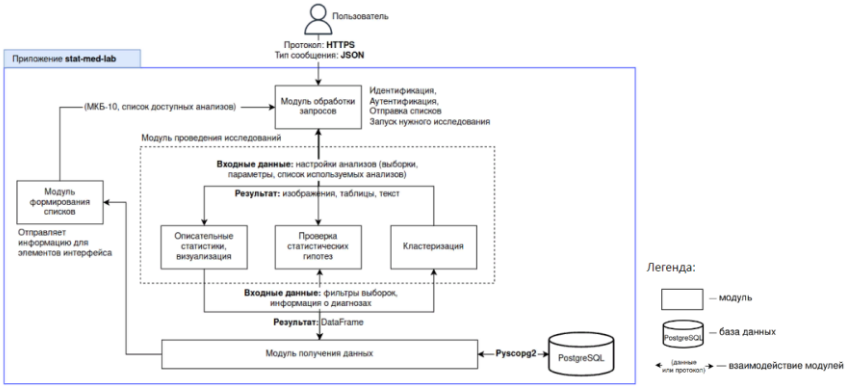

Рис. 1. Архитектура разрабатываемого приложения

Модули предобработки и импорта. Все исходные данные были заранее преобразованы из формата csv в формат parquet и заархивированы для снижения времени их считывания программой. Во время предобработки названия всех столбцов приводятся к единому виду за счет вручную созданных словарей для каждого датасета и также названия тестов общего и биохимического анализов крови (ОАК и БАК) приводятся к единому виду. На данный момент изменение названий тестов производится только для ОАК и БАК, потому что

они самые многочисленные и являются приоритетными для заказчика.

При импорте происходит загрузка предобработанных данных в БД результатов лабораторных исследований. Помимо этого, в базу данных загружается информация о заболеваниях МКБ [7] (об их древовидной структуре и полных названий) из файла формата csv.

Модуль статистического анализа

Далее описана реализация только трех основных методов из списка всех используемых в разрабатываемом приложении методов статистического анализа.

Иерархическая кластеризация

Входные данные: объект pandas DataFrame, где по строкам — направления, и по столбцам — возраст, пол (0 или 1), анализ1, анализ2, …, анализN.

Иерархическая кластеризация производится с помощью класса AgglomerativeClustering из библиотеки scikit-learn. Данная кластеризация проводится дважды: без указания количества кластеров для корректного построения дендрограммы и с указанием количества кластеров для корректного построения SPLOM (матрица диаграмм рассеяния "каждый параметр с каждым").

Выходные данные: метки, присвоенные каждому направлению; base64 изображения дендрограммы и SPLOM; описательная статистика для каждого кластера; список самых частых заболеваний у пациентов в каждом кластере.

На данный момент в приложении реализован только серверная часть иерархической кластеризации.

Кластеризация методом k-средних.

Входные данные: объект pandas DataFrame, где по строкам — направления, и по столбцам — возраст, пол (0 или 1), анализ1, анализ2, …, анализN; число кластеров, метрика расстояния.

Кластеризация k-средних производится с помощью класса kmeans из библиотеки pyclustering, так как данные библиотека и класс предоставляют возможность выбора метрики расстояния в отличии от похожего класса из библиотеки scikit-learn. Выбор используемой метрики осуществляется с помощью параметра metric при

инициализации объекта класса kmeans и выбирается из класса distance_metric (pyclustering.utils.metric).

Выходные данные: метки, присвоенные каждому направлению; base64 изображение SPLOM; описательная статистика для каждого кластера; список самых частых заболеваний у пациентов в каждом кластере.

Пример выхода метода представлен на рис. 2.

Рис. 2. Пример отображения результата работы метода *k*-средних в приложении

Дисперсионный анализ

Входные данные: минимум 2 объекта pandas DataFrame, где по строкам — направления, и по столбцам — результаты анализов; пороговое значение для нулевой гипотезы.

Однофакторный дисперсионный анализ производится с помощью функции f_oneway из библиотеки scipy, которая рассчитывает F-статистику и p-значение. Для расчета q-значения, которое обеспечивает средство контроля ложноположительных результатов, используется функция multipletests из библиотеки statsmodels.

Однофакторный дисперсионный анализ реализован только в серверной части приложения.

Выходные данные: вычисленная F-статистика теста; соответствующее p-значение из F распределения; булевские значения о принятии или непринятии нулевой гипотезы на основании p-значения; q-значение, вычисленное на основании p-значения; булевские значения о принятии или непринятии нулевой гипотезы на основании q-значения.

Пример выхода представлен на рис. 3.

	t-статистика	p-значение	отвержение h0 на основе p-значения	q-значение	отвержение h0 на основе q-значения
Гемоглобин	160.429	0.000	отвергается	0.000	отвергается
Лейкоциты	0.526	0.468	не отвергается	0.468	не отвергается
Эритроциты	34.190	0.000	отвергается	0.000	отвергается

Рис. 3. Пример отображения результата работы однофакторного дисперсионного анализа

Веб-клиент

Одним из самых главных требований заказчика к разрабатываемому проекту было развертывание разрабатываемого решения на хостинге.

Основной страницей веб-приложения является страница исследований. Данная страница разделена на 2 части: список выборок и история результатов исследований (рис. 4). Существуют возможности добавления, редактирования и удаления выборок. История результатов поддерживает отображение графиков, таблиц, текстовых описаний статистических инструментов. При нажатии на кнопку «Новое исследование» создается пустой шаблон для исследования, где пользователь может выбрать инструмент и настройки для выбранного инструмента.

Выборки могут на основе следующих фильтров: пол, возраст, диагнозы, район проживания и время сдачи анализа. Пользователь

может выбрать любую комбинацию диагнозов, в том числе и на разных уровнях. На рис. 5 приведено несколько примеров выборок.

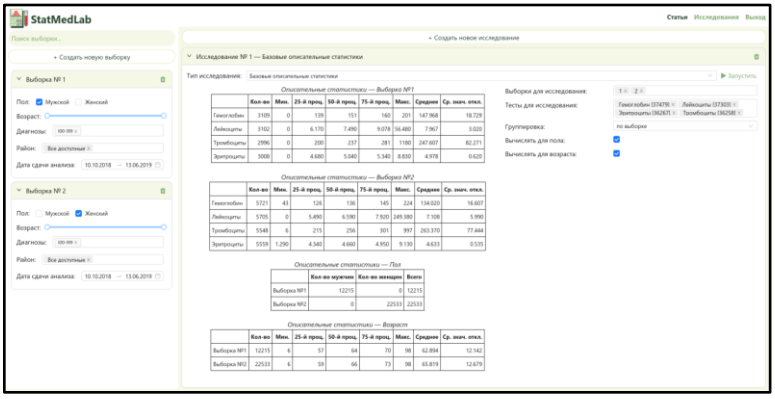

Рис. 4. Главная страница приложения

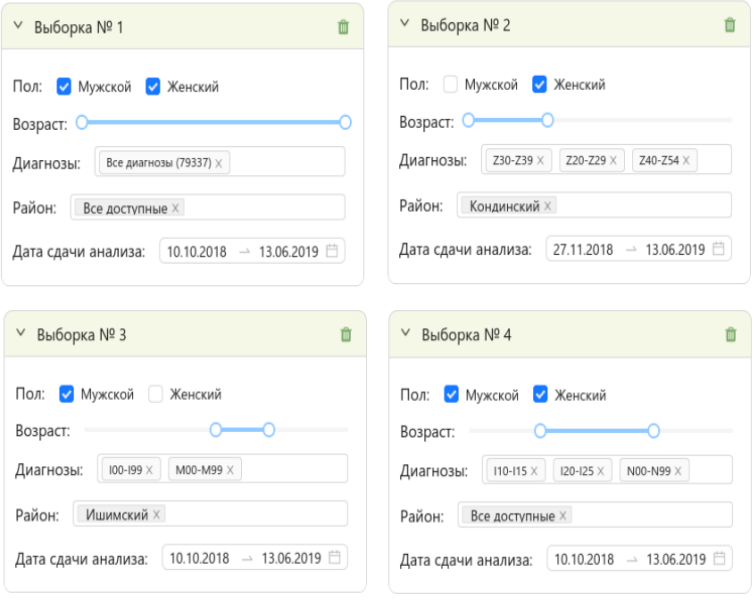

Рис. 5. Примеры разных выборок

Заключение. В процессе работы над проектом была изучена предметная область статистического анализа в медицине и были изучены следующие методы статистического анализа: иерархическая кластеризация, кластеризация k-средних, метод локального уровня выброса, метод главных компонент, t-критерии, медианный критерий, критерий χ-квадрат и дисперсионный анализ.

Реализованы модули предобработки и импорта данных в разработанную базу данных для хранения данных о результатах лабораторных анализов, а также реализованы серверная и клиентская части приложения значительной доли выбранных математических методов.

Главным преимуществом разработанного приложения перед аналогами является гибкая работы с выборками данных и интеграция с МКБ-10. Эта функциональность отсутствует у классических статистических пакетов. Простой интерфейс создания выборок позволяет быстро и легко формировать разные группы пациентов, сравнивать их между собой и оценивать статистические гипотезы об их сходствах или различиях.

В перспективе планируется доработать реализацию математических методов оценки статистических гипотез, улучшить клиентскую часть приложения, а также написать документацию серверного API приложения и руководство пользователя.

СПИСОК ЛИТЕРАТУРЫ

1. Armitage P., Berry G., Matthews J.N. Statistical methods in medical research. John Wiley & Sons, 2008. 816 p.
2. Petrie A., Sabin C. Medical statistics at a glance. John Wiley & Sons, 2019. 208 c.
3. Maimon O., Rokach L., editors. Data mining and knowledge discovery handbook. 2005. 1306 p.
4. Брюс П., Брюс Э., Гедек П. Практическая статистика для специалистов Data Science. Санкт-Петербург, 2021. 352 с.
5. Breunig M.M., Kriegel H.P., Ng R.T., Sander J. LOF: identifying density-based local outliers // In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000. 93-104 p.
6. Наркевич А.Н., Виноградов К.А. Выбор метода для статистического анализа медицинских данных и способа графического представления результатов // Социальные аспекты здоровья населения. 2019. №4 (68).
7. МКБ 10 — Международная классификация болезней 10-го пересмотра [Электронный ресурс] — URL: https://mkb-10.com/ (Дата обращения: 16.01.2023).