МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ ДЛЯ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА МЕДИЦИНСКИХ ДАННЫХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование
информационных систем
факультета КНиИТ
Шевцовой Варвары Михайловны

Научный руководитель
доцент, к. т. н. _____ М. В. Хамутова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ С. В. Миронов

Саратов 2026

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современных условиях здравоохранения возрастает потребность в оперативном анализе больших массивов медицинских данных. Такие данные часто многомерны и требуют применения комплекса статистических методов, что затрудняет их обработку без специализированных программных средств. Распространённые медицинские информационные системы ориентированы на учёт и стандартную отчётность, а встроенная статистика обычно ограничивается простыми сводными таблицами и базовыми графиками, чего недостаточно для проверки сложных гипотез и многомерного анализа. Специализированные же статистические продукты предоставляют широкий набор методов, но чаще всего являются коммерческими, требуют высокой квалификации и не обеспечивают единого удобного веб-интерфейса, из-за чего исследователю приходится комбинировать несколько инструментов или создавать временные решения.

Целью настоящей работы является разработка веб-приложения для многомерного статистического анализа медицинских данных, обеспечивающего единый интерактивный контур для основных этапов статистической обработки на основе табличных выгрузок из медицинских информационных систем и других источников.

Для достижения поставленной цели в работе решаются следующие задачи:

- обобщить теоретические основы статистического анализа медицинских данных, включая методы однофакторного и многомерного анализа;
- проанализировать существующие программные средства статистической обработки данных и выявить их ограничения в контексте работы с медицинскими выгрузками;
- определить требования к структуре входных данных и разработать модель представления медицинских индикаторов для унифицированной обработки табличных файлов;
- реализовать модули ключевых методов анализа (описательная статистика, когортный и «случай-контроль» анализ, проверка нормальности, корреляционный и регрессионный анализ, проверка гипотез, анализ главных компонент, кластерный и факторный анализ) на языке Python;
- интегрировать разработанные модули в интерактивное веб-приложение с поддержкой загрузки данных, визуализации результатов и формирования интерпретируемых выводов.

1 Теоретические основы статистических методов

В данном разделе рассматриваются статистические методы, наиболее часто используемые при анализе медицинских данных и положенные в основу разработанного веб-приложения.

1.1 Описательная статистика

Описательная статистика используется для первичного количественного и графического описания выборки медицинских данных. В качестве основных характеристик применяются меры центральной тенденции (среднее, медиана), меры вариабельности (дисперсия, стандартное отклонение, коэффициент вариации), а также показатели формы распределения (асимметрия, эксцесс). В контексте медицинских показателей эти характеристики позволяют оценить типичные значения и разброс индикаторов для выбранной группы пациентов или стран, а также выявить возможные выбросы и аномальные наблюдения.

1.2 Когортный анализ

Когортный анализ опирается на сравнение заранее выделенных групп наблюдений (когорт), сформированных по какому-либо признаку. Целью является оценка различий в целевом показателе между когортами. В качестве когорт могут выступать группы стран или пациентов с различающимися значениями категориального признака, а анализ проводится по непрерывному индикатору (например, ожидаемой продолжительности здоровой жизни). Для проверки значимости выявленных различий используются параметрические критерии.

1.3 Анализ «случай-контроль»

Исследование по схеме «случай-контроль» предполагает разделение наблюдений на две группы: «случаи» (обладающие исследуемым исходом) и «контроли» (отсутствие исхода). Далее проводится сравнение распределений второго признака между этими группами, что позволяет оценить возможную связь фактора и исхода. В рамках данной работы реализуется автоматизированное формирование групп по пороговому значению индикатора и последующая статистическая оценка различий между ними.

1.4 Регрессия и корреляционный анализ

Корреляционный анализ применяется для оценки силы и направления линейной зависимости между двумя количественными признаками. Наиболее распространённой мерой является коэффициент корреляции Пирсона; при наличии нарушений нормальности или существенных выбросов может использоваться ранговый коэффициент Спирмена. Регрессионный анализ дополняет корреляционный, позволяя построить регрессионную модель и количественно описать влияние одного показателя на другой, а также оценить качество аппроксимации через метрики средней квадратической ошибки и коэффициент детерминации. В контексте медицинских данных это даёт возможность изучать связь между различными индикаторами состояния здоровья и социально-экономическими факторами.

1.5 Критерий Смирнова-Колмогорова для проверки нормальности

Критерий Смирнова-Колмогорова относится к непараметрическим тестам согласия и используется для проверки гипотезы о соответствии эмпирического распределения выбранному теоретическому закону, чаще всего нормальному. Для медицинских данных это важно, поскольку многие последующие методы (например, параметрические критерии сравнения средних) предполагают нормальность распределения.

1.6 Графический метод Q-Q plot

Q-Q диаграмма представляет собой графический метод проверки согласия распределения данных с теоретическим законом. На диаграмме откладываются квантиль эмпирического распределения против квантилей теоретического распределения; отклонение точек от диагонали указывает на несоответствие модели. Построение Q-Q графика является наглядным дополнением к критериям проверки нормальности и позволяет пользователю визуально оценить поведение хвостов распределения и наличие выбросов.

1.7 Анализ главных компонент (PCA)

Анализ главных компонент (Principal Component Analysis, PCA) относится к методам снижения размерности и используется для перехода от множества коррелированных признаков к меньшему числу независимых компонент, объясняющих большую часть общей дисперсии. В задачах обработки медицинских

данных PCA позволяет выявить скрытые факторы, определяющие структуру индикаторов здоровья, а также визуализировать многомерные наблюдения в пространстве первых главных компонент.

1.8 T-тест Стьюдента для независимых выборок

T-тест Стьюдента для независимых выборок служит для проверки гипотезы о равенстве средних значений показателя в двух независимых группах. При выполнении стандартных предпосылок (нормальность распределения и близость дисперсий) этот тест позволяет сделать вывод о наличии статистически значимых различий между группами. В контексте разработанного веб-приложения t-тест используется как в когортном анализе, так и в исследовании «случай-контроль», обеспечивая автоматический расчёт статистики, p-значения и формирование краткого вывода.

1.9 Критерий хи-квадрат Пирсона

Критерий хи-квадрат Пирсона предназначен для анализа связи между двумя категориальными признаками на основе таблиц сопряжённости. Он позволяет проверить гипотезу о независимости признаков и оценить, существует ли статистически значимая ассоциация между ними. В медицинских данных данный критерий применяется, например, для проверки связи между наличием заболевания и принадлежностью к определённой группе, сформированной по демографическим или социальным характеристикам.

1.10 Кластерный анализ

Кластерный анализ представляет собой группу методов, направленных на разбиение множества объектов на кластеры таким образом, чтобы объекты внутри кластера были максимально похожи, а между кластерами — максимально различались. В медицинских исследованиях кластеризация позволяет выделять однородные группы стран или пациентов по совокупности показателей состояния здоровья, что способствует выявлению типичных профилей и потенциальных групп риска. В работе используется алгоритм k-средних, применяемый к нормированным значениям выбранных индикаторов.

1.11 Факторный анализ

Факторный анализ, близкий по идеологии к PCA, ориентирован на поиск скрытых факторов, обусловливающих наблюдаемые корреляции между пере-

менными. В отличие от PCA, факторный анализ вводит модель, в которой каждый наблюдаемый признак представляется линейной комбинацией меньшего числа факторов и уникальной компоненты. Применение факторного анализа к медицинским показателям позволяет группировать индикаторы в интерпретируемые блоки (например, связанные с демографической структурой, состоянием системы здравоохранения и т.п.), что облегчает интерпретацию результатов и формирование практических выводов по данным, обрабатываемым в разработанном веб-приложении.

2 Разработка информационной системы

В данном разделе описывается разработка веб-приложения для многомерного статистического анализа медицинских данных. Информационная система ориентирована на работу с табличными выгрузками показателей здоровья и обеспечивает загрузку файлов, предварительную обработку, применение статистических методов и визуализацию результатов в интерактивном режиме.

2.1 Цель и задачи программной реализации

Целью программной реализации является создание веб-приложения, предназначенного для многомерного статистического анализа агрегированных медицинских данных. Приложение должно обеспечивать пользователю удобный интерфейс для загрузки табличных данных, их предварительной обработки, выполнения основных методов статистического анализа и визуализации полученных результатов.

Основные задачи программной реализации включают в себя:

- реализацию средств загрузки и предварительной обработки входных данных, включая контроль структуры файла, приведение типов и базовую очистку;
- разработку модуля описательной статистики для получения ключевых агрегированных характеристик показателей и детального анализа распределения выбранного показателя;
- построение модуля когортного анализа для сопоставления двух когорт по количественным показателям;
- построение модуля исследования по схеме «случай-контроль» для сравнения групп, сформированных на основе порогового значения индикатора;
- реализацию модуля корреляционного и регрессионного анализа для оценки связей между показателями и построения линейных моделей;
- реализацию модуля проверки нормальности распределения с использованием нескольких критериев и графического анализа;
- реализацию модуля анализа главных компонент для снижения размерности и выявления структуры взаимосвязей между показателями;
- реализацию модуля проверки статистических гипотез на основе Т-теста Стьюдента и критерия хи-квадрат;
- реализацию модуля кластерного анализа для выделения однородных групп

стран по совокупности показателей;

— реализацию модуля факторного анализа для выявления латентных факторов и интерпретации структуры многомерных данных;
— обеспечение возможности визуального представления результатов анализа в виде таблиц и графиков, а также экспорта полученных статистических сводок;
— поддержку интуитивно понятного пользовательского интерфейса, позволяющего выполнять перечисленные аналитические операции в интерактивном режиме.

2.2 Архитектура и технический стек

Выбор технического стека обусловлен необходимостью поддержки многомерного статистического анализа, работы с табличными медицинскими данными и поэтапного расширения функциональности без изменения базовой структуры системы. В качестве основной платформы выбран язык Python и связанный с ним стек библиотек, поскольку он позволяет объединить в едином технологическом контуре обработку данных, статистические вычисления, визуализацию и пользовательский веб-интерфейс.

Веб-приложение реализовано в виде одностраничного аналитического интерфейса на базе фреймворка Streamlit. Такой подход позволяет описывать логику взаимодействия с пользователем и вычислительные процедуры в одном коде, упрощает развёртывание системы и снижает порог входа для конечного пользователя.

Архитектурно система разделена на три взаимосвязанных уровня. Уровень данных отвечает за загрузку файлов CSV/XLS/XLSX, их преобразование к унифицированной структуре и хранение в виде объектов `pandas.DataFrame`. Уровень бизнес-логики включает аналитические модули, реализующие методы статистического анализа. Уровень представления образуют интерфейсные компоненты Streamlit, обеспечивающие выбор параметров, запуск расчётов, просмотр таблиц и графиков, а также экспорт результатов.

В качестве основных библиотек используются `pandas` для табличной обработки и работы с `DataFrame`, `numpy` для базовых численных операций, `matplotlib` и `seaborn` для построения графиков и диаграмм, инструменты `scipy.stats` для реализации статистических критериев (проверка нормальности, Т-тест и другие), а также компоненты `scikit-learn` для линейной ре-

грессии, анализа взаимосвязей, анализа главных компонент, кластерного и факторного анализа. Такое сочетание библиотек позволяет реализовать в рамках единого расширяемого веб-приложения все необходимые модули.

2.3 Реализация основных модулей

Реализация аналитической части приложения организована модульно: каждому классу статистических методов соответствует отдельный программный модуль, интегрированный в общий интерфейс.

1. Модуль описательной статистики предоставляет пользователю выбор показателя и, при необходимости, категориальной переменной. На основе выбранных данных вычисляются основные числовые характеристики, формируется сводная таблица и строятся гистограммы распределения, а также дополнительные графики, такие как диаграмма размаха и ранжированный столбчатый график.
2. Модуль когортного анализа ориентирован на сравнение двух категорий выбранного признака. В рамках модуля выполняется формирование двух групп наблюдений, вычисление их описательных характеристик и применение t-теста Стьюдента для независимых выборок. Результаты представляются в виде таблицы с основными метриками и графика boxplot с наложением индивидуальных точек.
3. Модуль анализа по схеме «случай-контроль» реализует автоматический подбор порогового значения по выбранному индикатору, формирование групп «случаи» и «контроли», а также сравнение вторичного количественного показателя между этими группами. Вычисляются описательные характеристики, t-статистика и p-значение, строится графическая визуализация распределений.
4. Модуль корреляционного и регрессионного анализа принимает два количественных показателя и, при необходимости, ограничение по категории. В модуле рассчитываются коэффициенты корреляции Пирсона и Спирмена, строится линейная регрессионная модель с оценкой качества по среднеквадратичной ошибке и коэффициенту детерминации. Графически результаты иллюстрируются диаграммой рассеяния с регрессионной прямой, матрицей корреляции и графиком остатков.
5. Модуль проверки нормальности объединяет числовые и графические подходы. Для выбранного показателя вычисляются статистики критериев

проверки нормальности и соответствующие p-значения, дополнительно строится Q-Q диаграмма, что позволяет пользователю сопоставить численные выводы с визуальной оценкой.

6. Модуль анализа главных компонент реализует снижение размерности для набора выбранных индикаторов. Вычисляются доли объяснённой дисперсии по компонентам, формируется scree-плот, а также матрица нагрузок для интерпретации компонент. При использовании двух первых компонент объекты визуализируются на плоскости, что позволяет выявлять группы и аномальные наблюдения.
7. Модуль проверки статистических гипотез предназначен для выбора пользователем типа теста (например, t-теста для независимых выборок или критерия хи-квадрат), задания сравниваемых групп и автоматического расчёта соответствующих статистик и p-значений. Результаты представляются в виде компактной сводной таблицы и сопровождаются текстовой интерпретацией, указывающей на принятие или отклонение нулевой гипотезы при заданном уровне значимости.
8. Модуль кластерного анализа обеспечивает автоматическое разбиение наблюдений на заданное пользователем число кластеров на основе нормированных значений выбранных индикаторов, для чего в качестве базового алгоритма применяется метод k-средних; в результате формируются таблицы с характеристиками кластеров и двумерная визуализация в пространстве главных компонент с цветовой маркировкой кластеров.
9. Модуль факторного анализа предназначен для выделения скрытых факторов, лежащих в основе взаимосвязей между медицинскими показателями; в рамках модуля вычисляются факторные нагрузки и общности, формируется тепловая карта нагрузок и таблица дисперсии, объясняемой факторами, что обеспечивает пользователю инструмент для интерпретации структуры данных и построения укрупнённых индикаторов. Все модули объединены в единую навигационную панель приложения, что позволяет пользователю в рамках одной сессии загружать данные, переключаться между методами анализа и последовательно интерпретировать полученные результаты.

3 Апробация и результаты

В целях апробации разработанного веб-приложения был использован открытый набор данных Всемирной организации здравоохранения, содержащий показатели, характеризующие состояние здоровья населения и параметры системы здравоохранения для различных стран и периодов наблюдения. Данные были приведены к унифицированному табличному виду, осуществлена проверка полноты и корректности значений, выполнено кодирование категориальных признаков и фильтрация индикаторов, пригодных для статистического анализа.

На первом этапе с помощью модуля описательной статистики были получены сводные характеристики по ключевым медицинским показателям, построены гистограммы распределений и диаграммы размаха. Это позволило выявить особенности вариабельности показателей, наличие выбросов и различий между группами стран, а также определить индикаторы, требующие более детального анализа.

Далее были применены модули когортного анализа и анализа по схеме «случай-контроль» для сравнения групп, сформированных по демографическим и социально-экономическим признакам. На основе t-теста Стьюдента и критериев проверки нормальности были выявлены статистически значимые различия в целевых показателях между когортами, что подтверждает практическую применимость реализованных методов к реальным медицинским данным.

Модуль корреляционного и регрессионного анализа был использован для изучения связей между показателями здоровья и вспомогательными индикаторами. Расчёт коэффициентов корреляции, построение регрессионных моделей и анализ графиков остатков позволили выявить ряд устойчивых зависимостей, а также продемонстрировать пользователю ограничения линейных моделей при наличии нелинейных эффектов и выбросов.

С помощью модуля проверки нормальности и построения Q-Q диаграмм была оценена корректность применения параметрических критериев к выбранным показателям. Для части индикаторов выявлено существенное отклонение от нормального распределения, что обосновывает необходимость использования непараметрических подходов или предварительных преобразований данных при дальнейших исследованиях.

Модуль проверки статистических гипотез обеспечивает выбор типа теста, задание сравниваемых групп и автоматический расчёт соответствующих

статистик и p-значений; результаты сопровождаются краткой текстовой интерпретацией, указывающей на принятие или отклонение нулевой гипотезы при заданном уровне значимости.

Модуль анализа главных компонент позволил сократить размерность исходного пространства признаков и выделить несколько компонент, объясняющих значительную долю общей дисперсии показателей. Визуализация стран в пространстве первых главных компонент выявила группы с близкими профилями здоровья и позволила обнаружить аномальные наблюдения. На основе тех же признаков был выполнен кластерный анализ методом k-средних; полученные кластеры интерпретированы как группы стран со схожими значениями медицинских индикаторов, а дополнительное применение факторного анализа позволило сгруппировать показатели в интерпретируемые факторы и уточнить структуру связей между индикаторами.

Ниже представлены примеры интерфейса и работы приложения (рисунки 3.1 и 3.2).

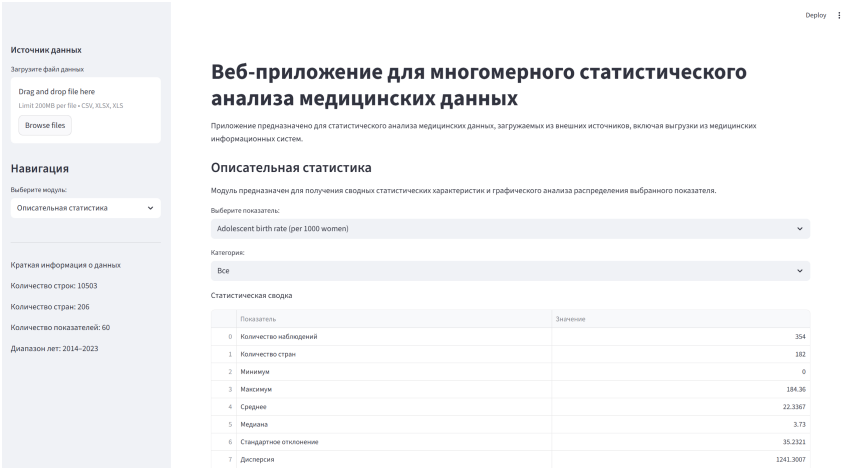

Рисунок 3.1 – Пример работы приложения

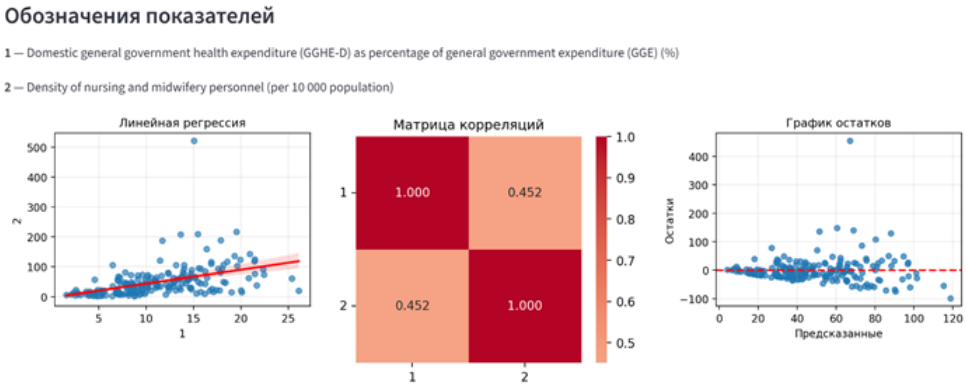

Рисунок 3.2 – Пример работы приложения

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы поставленная цель исследования, заключающаяся в разработке веб-приложения для многомерного статистического анализа медицинских данных, достигнута. Были решены следующие задачи:

- обобщены теоретические основы статистического анализа медицинских данных, включая методы однофакторного и многомерного анализа;
- проанализированы существующие программные средства статистической обработки данных и выявлены их ограничения в контексте работы с медицинскими выгрузками;
- определены требования к структуре входных данных и разработана модель представления медицинских индикаторов, обеспечивающая унифицированную обработку табличных файлов;
- реализованы модули описательной статистики, когортного и «случай-контроль» анализа, проверки нормальности распределений, корреляционного и регрессионного анализа, проверки статистических гипотез, анализа главных компонент, кластерного и факторного анализа;
- разработанные модули интегрированы в архитектуру интерактивного веб-приложения с поддержкой загрузки данных, визуализации результатов и формирования интерпретируемых выводов.

Разработанное приложение обеспечивает обработку данных заданной структуры, загрузку пользовательских файлов, визуализацию и получение интерпретируемых статистических выводов. Примеры с использованием реальных международных показателей показали, что система позволяет выявлять закономерности в многомерных данных и связи между ключевыми индикаторами.

Научная новизна работы состоит в реализации модульной веб-ориентированной архитектуры, объединяющей в одном приложении широкий спектр методов многомерного статистического анализа и ориентированной на работу с агрегированными выгрузками. Практическая значимость определяется возможностью использования созданного программного комплекса в образовательных и исследовательских задачах. Результаты исследования представлены на XXVIII Всероссийской научно-практической конференции молодых учёных с международным участием «Россия молодая», по теме магистерской диссертации подготовлена статья для публикации в сборнике материалов конференции, подлежащем регистрации в РИНЦ.