УДК 004

РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА МЕДИЦИНСКИХ ДАННЫХ НА ОСНОВЕ МЕТОДОВ СНИЖЕНИЯ РАЗМЕРНОСТИ И ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Шевцова В.М., студентка гр. 273, 2 курс

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

Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского

Г. Саратов

Введение

В условиях цифровизации здравоохранения и развития биомедицинских исследований наблюдается значительный рост объёмов медицинских данных. Эти данные включают результаты лабораторных анализов, клинические показатели, эпидемиологическую информацию и другие параметры, требующие систематизации и анализа. Эффективное использование таких данных невозможно без применения методов математической статистики.

Статистический анализ играет ключевую роль в медицине, обеспечивая возможность выявления закономерностей, оценки достоверности результатов и принятия обоснованных решений. Например, методы описательной статистики используются для анализа клинических показателей, тогда как корреляционный анализ позволяет выявлять взаимосвязи между биомаркерами и исходами лечения, при оценке эффективности лекарственных препаратов применяется проверка статистических гипотез, а методы многомерного анализа используются в задачах диагностики и стратификации пациентов. Однако практическое применение статистических методов часто затруднено сложностью существующих инструментов и требованием специальных знаний. В связи с этим возникает необходимость в разработке доступных и удобных систем анализа данных.

Целью данной работы является разработка веб-приложения, автоматизирующего статистический анализ медицинских данных и упрощающего взаимодействие рядового пользователя с аналитическими инструментами.

Обзор существующих решений

Существующие инструменты для статистического анализа можно условно разделить на программные пакеты, решающие определённые задачи, и языки программирования. В медицинской статистике широкое применение

нашли языки R и MATLAB, хорошо подходящие для анализа данных. R ориентирован на статистические вычисления и имеет большое количество библиотек, но его использование требует соответствующей квалификации. MATLAB эффективен в задачах численного анализа, но является коммерческим продуктом, что ограничивает его доступность для ряда организаций.

В научных исследованиях активно применяется Python, являющийся универсальным языком программирования. Но несмотря на простоту синтаксиса, наличие разнообразных библиотек и широкую поддержку со стороны сообщества, в том числе научного, как и с любым другим языком программирования его использование требует специальных навыков.

Постановка задачи

Для достижения цели работы необходимо разработать веб-приложение, обеспечивающее полный цикл статистической обработки медицинских данных, включая загрузку данных, их предварительную обработку, анализ и визуализацию результатов. На этапе подготовки данных должна осуществляться их очистка, так как медицинские данные часто содержат пропуски и статистические выбросы, обработка которых необходима для обеспечения достоверности последующего анализа. Аналитическая часть должна базироваться на методах математической статистики, позволяя пользователю выполнять статистический анализ с формализованной интерпретацией результатов.

Выбор инструментов и библиотек Python

Для реализации приложения выбраны язык Python и набор специализированных библиотек: pandas для работы с табличными данными (включая загрузку, фильтрацию и преобразование), NumPy для выполнения численных операций и работы с массивами, scipy.stats для реализации статистических тестов и проверки гипотез. Методы машинного обучения, включая анализ главных компонент, реализованы с помощью библиотеки scikit-learn. Данные визуализируются с помощью matplotlib и seaborn. Веб-интерфейс построен на Streamlit, что позволяет создавать интерактивные приложения без значительных трудозатрат на разработку пользовательского интерфейса.

Архитектура приложения

В основе архитектуры приложения лежит модульный принцип, обеспечивающий расширяемость и удобство сопровождения. Каждый этап анализа представляет собой отдельный функциональный модуль. Пользователь взаимодействует с системой через веб-интерфейс, в котором доступны загрузка данных и выбор методов анализа. После загрузки данные могут использоваться любыми модулями, что исключает повторную обработку и повышает эффективность системы.

Реализация основных функций и демонстрация результатов

Функционал веб-приложения направлен на комплексное решение главных задач статистического анализа в медицине. Отдельное внимание уделено интеграции теоретических методов математической статистики в прикладную программную среду, что позволяет пользователю, не имеющему углубленных знаний статистического аппарата, получать корректные и интерпретируемые результаты.

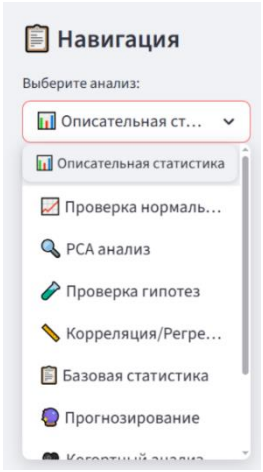

Рисунок 1 – Расположение модулей в пользовательском интерфейсе при выборе метода анализа

На первом этапе выполняется загрузка и предварительная обработка данных. Медицинские данные, как правило, содержат пропуски, неоднородны по структуре и часто включают выбросы. Для исключения пропущенных значений используется метод dropna, что гарантирует корректность дальнейших вычислений.

После подготовки данных следует этап описательной статистики, играющий важную роль в первичном анализе выборки. Рассчитываются основные числовые характеристики, отражающие центральную тенденцию и вариабельность. Полученные показатели позволяют оценить распределение значений, выявить асимметрию и определить степень разброса. Такие статистические характеристики, как среднее арифметическое, медиана, стандартное отклонение, дисперсия, коэффициент вариации, асимметрия и эксцесс, рассчитываются с помощью встроенных методов pandas.Series, а также функций scipy.stats.skew и scipy.stats.kurtosis, позволяющих оценить отклонение распределения от нормального.

Для визуализации распределения используется гистограмма, строящаяся с помощью matplotlib.pyplot.hist. С помощью axvline дополнительно отображаются линии среднего и медианы, что даёт наглядное представление о центральной тенденции и разбросе данных. Итоговые результаты формируются в виде структурированного словаря и отображаются в интерфейсе посредством st.json.

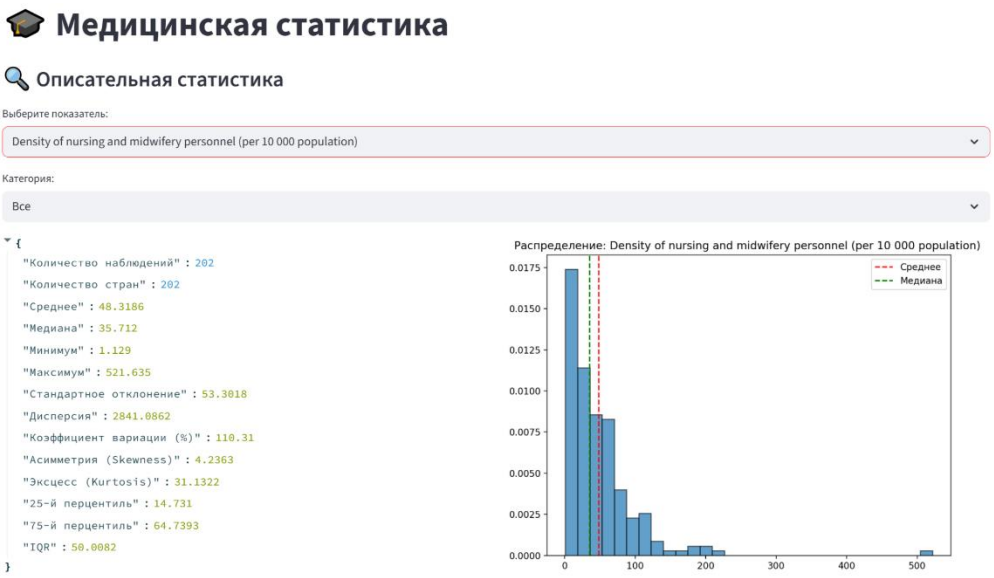

Рисунок 2 – Модуль описательной статистики

Следующим этапом идёт анализ распределений. Проверка соответствия данных нормальному распределению имеет принципиальное значение для выбора дальнейших методов анализа, так как многие статистические критерии основаны на предположении о нормальности. В модуле проверки нормальности распределения реализованы три статистических критерия: Шапиро–Уилка, Колмогорова–Смирнова и Д’Агостино, каждый из которых реализован с использованием соответствующих функций из библиотеки scipy.stats. Критерий Шапиро–Уилка вызывается через shapiro, что позволяет оценить нормальность выборки малого и среднего объема. Критерий Колмогорова–Смирнова реализуется функцией kstest, в которой в качестве эталонного распределения задается нормальное, параметры которого оцениваются по выборке. Критерий Д’Агостино, предназначенный для проверки нормальности на основе асимметрии и эксцесса, вызывается через normaltest.

Для визуальной оценки соответствия распределения нормальной модели применяется Q-Q plot, построенный с помощью stats.probplot из scipy.stats, который отображает квантили выборки относительно теоретических квантилей нормального распределения. Визуализация сопровождается интерпретацией результатов в виде метрик с цветовой индикацией, что позволяет пользователю быстро принять решение о применимости параметрических методов в дальнейшем анализе.

Анализ взаимосвязей между переменными осуществляется с использованием методов корреляционного анализа. На этом этапе определяются степень и направление зависимости между показателями. В медицинских исследованиях это может использоваться для выявления взаимосвязей между биомаркерами, параметрами состояния пациента и результатами лечения.

Инструментом для более глубокого анализа зависимостей является регрессионный анализ, позволяющий моделировать влияние одной или нескольких переменных на исследуемый показатель. Регрессионные модели не только описывают наблюдаемые зависимости, но и служат основой для прогнозирования.

Вычисление коэффициентов корреляции Пирсона и Спирмена в модуле корреляционного и регрессионного анализа реализовано с помощью функций pearsonr и spearmanr из библиотеки scipy.stats. Линейная регрессия строится с помощью seaborn.regplot, который автоматически вычисляет и отображает линию регрессии и коэффициент детерминации. Визуализация дополняется тепловой картой корреляционной матрицы seaborn.heatmap, значения коэффициентов корреляции отображаются с точностью до трёх знаков после запятой. Для обеспечения парного соответствия наблюдений выполняется объединение данных по странам с помощью pd.merge. Визуальные и числовые результаты представляются в двухколоночном формате, обеспечивающем наглядное сравнение линейной и ранговой корреляции и позволяющем оценить силу и направление связи между двумя медицинскими показателями.

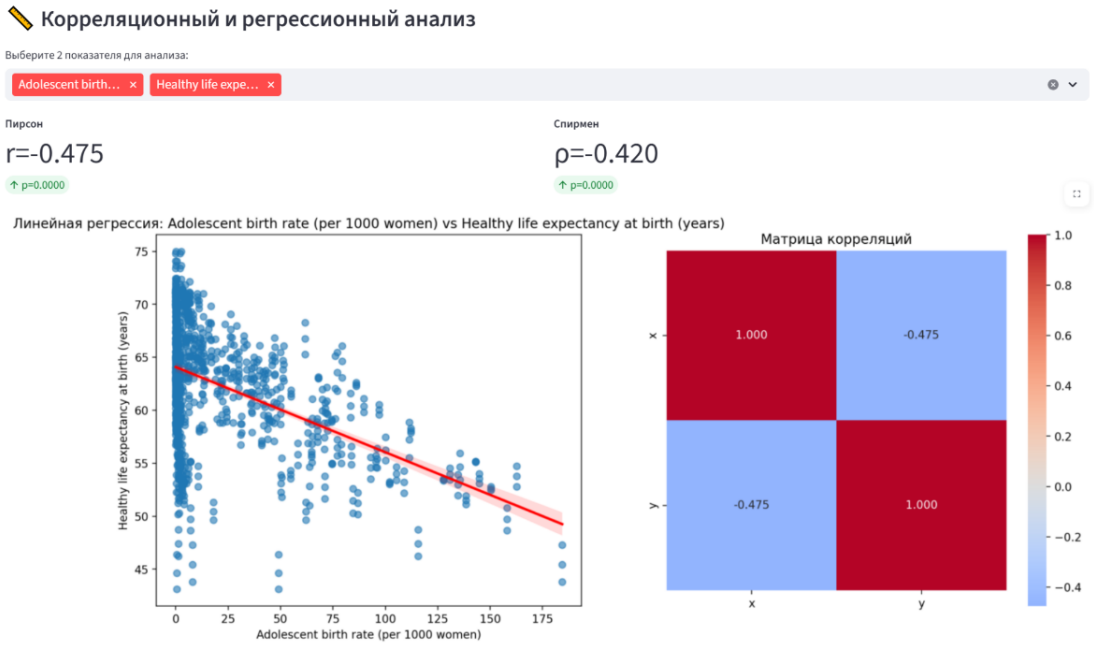

Рисунок 3 – Модуль корреляционного и регрессионного анализа

Отдельный блок посвящён процедурам проверки статистических гипотез. В приложении реализованы методы сравнения выборок и оценки статистической значимости различий. Результаты тестов дополняются интерпретацией, основанной на сравнении уровня значимости с заданным пороговым значением. Модуль проверки статистических гипотез содержит t-тест Стьюдента для независимых выборок, использующий функции ttest_ind из scipy.stats. Обрабатываемые данные предварительно фильтруются по выбранному показателю и двум категориям, после чего извлекаются значения для каждой группы с удалением пропущенных наблюдений. Для визуализации

межгрупповых различий применяется boxplot из seaborn — данный график даёт представление о распределении, медиане, выбросах и разбросе значений. Результаты теста отображаются в виде метрик: t-статистика и p-значение, сопровождаемые интерпретацией на основе уровня значимости 0.05.

В рамках многомерного анализа реализован метод главных компонент, позволяющий работать с высокоразмерными данными. Данный метод обеспечивает снижение размерности пространства признаков за счёт выделения наиболее информативных компонент. Это позволяет выявлять скрытые закономерности в данных и упрощает их дальнейший анализ. Модуль реализован с использованием класса PCA из библиотеки scikit-learn, а также функций StandardScaler для предварительной стандартизации данных. Перед выполнением PCA данные преобразуются в сводную таблицу с помощью pivot_table, где строки соответствуют странам, а столбцы — выбранным показателям. После стандартизации с помощью fit_transform выполняется декомпозиция данных на главные компоненты. Объясненная дисперсия по каждой компоненте вычисляется через атрибут explained_variance_ratio, а накопленная дисперсия рассчитывается с помощью np.cumsum.

Визуализация результатов осуществляется через Scree Plot, построенный с помощью matplotlib.pyplot.plot, где также отображается пороговая линия, соответствующая среднему вкладу компонент. Вклад исходных признаков в компоненты вычисляется как произведение матрицы компонентов на корень из объясненной дисперсии и представляется в виде таблицы с помощью st.dataframe, что позволяет интерпретировать физический смысл каждой главной компоненты.

Тестирование и валидация

Тестирование веб-приложения проводилось на медицинских данных с различными характеристиками, в частности, использовался набор данных World Health Statistics, содержащий статистические медицинские показатели по странам за 2024 год. Проверялись и оценивались корректность вычислений, устойчивость к пропущенным значениям и адекватность визуализации. Результаты тестирования показали, что заявленные статистические методы реализованы корректно и приложение может быть использовано для анализа данных на практике.

Заключение

В результате выполненной работы разработано веб-приложение, автоматизирующее статистический анализ медицинских данных. Реализованные методы позволяют проводить комплексный анализ, включающий проверку гипотез, анализ зависимостей и многомерную обработку. Результаты анализа по каждому модулю сопровождаются соответствующим графическим представлением, что крайне важно для интерпретации: визуализация позволяет человеку выявить закономерности,

трудно обнаруживаемые при анализе числовых таблиц, повышает наглядность и облегчает восприятие результатов.

Разработанное решение можно применять в научных исследованиях, клинической практике и учебном процессе. Дальнейшее развитие предполагает расширение функциональных возможностей и внедрение методов машинного обучения.

Список литературы:

1. Токмачев, М.С. Математическая статистика в медицине [Текст] / Токмачев, М.С. и Медик, В.А. — Санкт-Петербург : Издательство «Лань», 2021. — Т. 1. — ISBN: 978-5-8114-6325-4.
2. Заварукин, А.С. Статистический анализ данных медицинских исследований [Текст] / Заварукин, А.С., Гайдаков, Г.А. и Борисов, Д.Н. // Известия Российской военно-медицинской академии. — 2022. — Т. 41, № S2. — С. 160–162.
3. Кайдаракова, Е.А. Кластерный анализ данных медицинской статистики [Текст] // Инженерные технологии: традиции, инновации, векторы развития. — Абакан : Хакасский государственный университет имени Н. Ф. Катанова. — 2024. — С. 29–30.
4. Курбанова, Н.Г. Применение факторного анализа в медицине [Текст] // Актуальные проблемы современного образования. Симметрии: теоретический и методический аспекты. — Астрахань : Астраханский государственный университет имени В. Н. Татищева. — 2024. — С. 81–85.
5. Чехлистова, Ю.А. Проверка гипотезы о распределении. Критерий Пирсона [Текст] / Чехлистова, Ю.А., Ельшин, В.Р. и Глазкова, М.Ю. // International Journal of Professional Science. — 2024. — Т. 11, № 2. — С. 65–73.
6. Проектирование и программная реализация комплекса анализа данных медико-статистической информации [Текст] / Есауленко, И.Э., Артемов, М.А., Кириченко, Д.О., Рудалев, В.Г. и Серженко, Н.П. // Системный анализ и управление в биомедицинских системах. — 2012. — Т. 11, № 2. — С. 301–305.
7. Черномордов, С.Б. Подход к моделированию и анализу данных на основе развития алгоритмов машинного обучения [Текст] / Черномордов, С.Б. и Опенкин, Д.Ю. // Continuum. Математика. Информатика. Образование. — 2021. — Т. 2, № 22. — С. 98–108.
8. Зубкова, Н.С. Применение метода главных компонент для анализа факторов появления сахарного диабета [Текст] // Летняя школа молодых ученых ЛГТУ. Сборник трудов научно-практической конференции студентов Липецкого государственного технического университета. —

Липецк : Липецкий государственный технический университет. — 2024. — С. 76–80.
9. McKinney, Wes. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter [Text]. — 3 ed. — [S. l.] : O'Reilly Media, 2022. — ISBN: 9781098104030.
10. Glantz, S.A. Primer of Applied Regression Analysis of Variance [Text] / Glantz, S.A. and Slinker, B.K. — 4 ed. — New York : McGraw-Hill Education, 2021. — ISBN: 978-1260455341.
11. Токмачев, М.С. Математическая статистика в медицине [Текст] / Токмачев, М.С. и Медик, В.А. — Санкт-Петербург : Издательство «Лань», 2021. — Т. 1. — ISBN: 978-5-8114-6325-4.