|  |  |  |
| --- | --- | --- |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  РОССИЙСКОЙ ФЕДЕРАЦИИ  Федеральное государственное бюджетное  образовательное учреждение высшего образования  «Пермский государственный национальный  исследовательский университет» | | |
|  | Кафедра прикладной математики и информатики | |
| **СОЗДАНИЕ ПРИЛОЖЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ВОЗНИКНОВЕНИЯ ОСЛОЖНЕНИЙ ПРИ ПЕРИТОНИТЕ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ.** | | |
| *Курсовая работа* | | |
|  | | |
|  | | |
|  | | Работу выполнил студент  группы ПМИ-4 3 курса физико-математического института  Сухой Д.В. |
|  | | Научный руководитель:  директор физико-математического института ПГНИУ  Барулина М.А.  «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2024 г. |
|  | |  |
| Пермь 2024 | | |

Оглавление

[Введение 3](#_gjdgxs)

[Глава 1. Обзор подходов 5](#_30j0zll)

[Глава 2. Исследование и обработка данных 7](#_1fob9te)

[Глава 3. Построение моделей 11](#_3znysh7)

[Глава 4. Развертывание модели 15](#_2et92p0)

[Глава 5. Заключение 16](#_tyjcwt)

[Глава 6. Библиографический список 17](#_3dy6vkm)

# Введение

Целью данного исследования является разработка системы прогнозирования осложнений, возникающих при перитоните, с использованием алгоритмов машинного обучения.

Перитонит (Peritonitis) — это воспаление брюшины (тонкой плёнки, которая покрывает внутреннюю поверхность брюшной полости и органы, расположенные в ней). Симптомы перитонита: боль в животе, температура выше 38 °C, вздутие живота, приступы тошноты с рвотой. Без лечения человек погибает в течение нескольких дней. Поэтому при появлении симптомов перитонита нужно немедленно обратиться за медицинской помощью.

Перитонит остается одной из актуальных проблем из-за его высокой летальности при несвоевременном или недостаточном лечении, а также возникновении послеоперационных осложнений, которые существенно осложняют реабилитацию и восстановление пациента. Поэтому модели, позволяющие прогнозировать возможное возникновение осложнений на основе данных, полученных при госпитализации пациента, несомненно, смогут помочь скорректировать схему лечения и повысить качество реабилитационного периода.

Несмотря на достижения современной медицины, раннее выявление и профилактика осложнений остаются сложной задачей из-за индивидуальных особенностей пациента и сложного взаимодействия различных факторов риска.

Для решения этих проблем мы предлагаем использовать алгоритмы машинного обучения для выявления закономерностей в большом наборе данных клинических переменных, связанных с пациентами с перитонитом. Наш подход имеет ряд преимуществ по сравнению с традиционными диагностическими инструментами, включая повышенную точность, масштабируемость и адаптивность к изменяющимся группам пациентов.

В этой работе описываются наши усилия по сбору и анализу полного набора клинических данных о пациентах с перитонитом. Мы также описываем шаги, предпринятые для предварительной обработки и подготовки данных к моделированию.

Затем мы сравниваем производительность трех популярных алгоритмов машинного обучения — машин с опорными векторами (SVM), градиентный бустинг и k-ближайших соседей (KNN) — чтобы определить наиболее эффективный алгоритм прогнозирования осложнений перитонита. Для оценки производительности модели мы рассчитываем стандартные показатели оценки (метрики), такие как precision, recall, ROC-AUC и F1. На основе этих метрик мы выбираем наиболее эффективный алгоритм и настраиваем его параметры с помощью перекрестной проверки, поиска по сетке и настройки гиперпараметров.

Наконец, мы демонстрируем практическую полезность нашей прогностической модели, реализуя ее в виде веб-приложения с использованием библиотеки Streamlit, позволяющего врачам вводить информацию о пациентах и получать мгновенную обратную связь относительно прогнозируемой вероятности развития осложнений. Это внедрение представляет собой значительный шаг на пути улучшения диагностики перитонита и результатов лечения.

Подводя итог, в этой работе представлен новый подход к прогнозированию осложнений перитонита с использованием алгоритмов машинного обучения. Благодаря анализу и тщательной реализации мы показали, что наша прогностическая модель обладает как теоретическими, так и практическими преимуществами по сравнению с традиционными диагностическими инструментами. В дальнейшем расширение объема набора данных, включение дополнительных клинических переменных и изучение альтернативных подходов к машинному обучению

# Глава 1. Обзор подходов

Хотя традиционные статистические методы уже давно служат надежными инструментами для анализа медицинских данных, последние разработки в области машинного обучения (ML) показали большие перспективы в решении все более сложных задач здравоохранения. Ниже мы рассмотрим различия между ML и традиционными статистическими подходами, выделив причины, по которым ML может быть предпочтительнее для прогнозирования осложнений перитонита.

1. Обработка сложных взаимосвязей: в отличие от традиционных статистических методов, которые в первую очередь фокусируются на линейных зависимостях между переменными, алгоритмы ML превосходно улавливают нелинейные зависимости и взаимодействия более высокого порядка. Эта возможность позволяет моделям ML лучше представлять истинную природу взаимосвязей в медицинских данных, потенциально давая более точные прогнозы.

2. Большие наборы данных: Традиционные статистические методы с трудом справляются с чрезвычайно большими наборами данных из-за вычислительных ограничений и предположений о независимости между наблюдениями. Напротив, алгоритмы ML могут эффективно обрабатывать огромные объемы данных, в полной мере используя преимущества современных аппаратных ускорителей (например, графических процессоров) и парадигм распределенных вычислений.

3. Непараметрический характер: Большинство традиционных статистических методов накладывают строгие параметрические допущения на вид распределения данных, что ограничивает их гибкость и применимость. Однако алгоритмы ML, как правило, являются непараметрическими, требующими минимальных предположений о распределении, но при этом обеспечивающими высокую производительность.

4. Универсальность для разных типов задач: Различные алгоритмы ML подходят для различных задач прогнозирования, будь то классификация, регрессия или кластеризация. Традиционные статистические методы обычно фокусируются на одном конкретном аспекте, что ограничивает их область использования.

5. Активное обучение и онлайн-тренинг: алгоритмы ML могут динамически корректировать прогнозы на основе поступающих данных с помощью схем активного обучения и онлайн-тренинга. Постоянное обновление поддерживает модели ML актуальными и реагирующими на возникающие тенденции, в то время как традиционные статистические методы полагаются исключительно на статический автономный анализ.

Учитывая эти сравнения, ML обладает неоспоримыми преимуществами перед традиционными статистическими методами прогнозирования осложнений перитонита. Его способность обрабатывать сложные взаимосвязи, большие массивы данных, адаптироваться к различным типам задач делает его мощным инструментом для раскрытия сложных взаимосвязей в обширных медицинских данных.

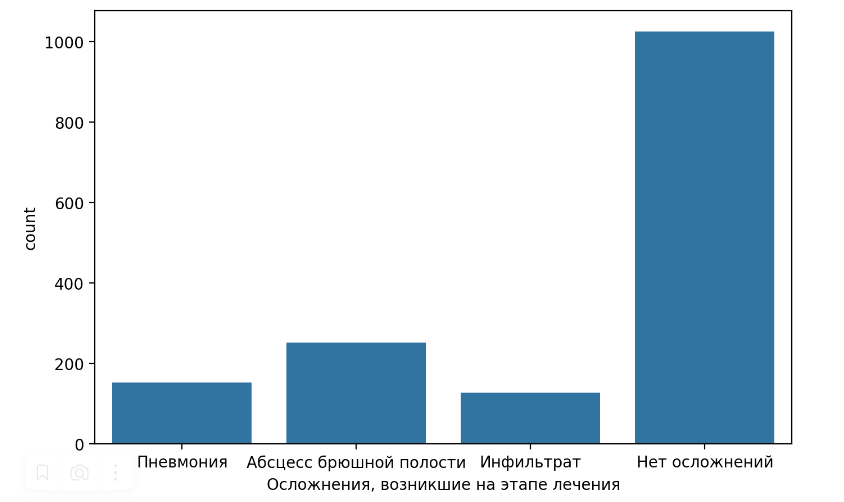
Предоставляя подробную информацию о существующих методах и сравнивая их с подходами машинного обучения, мы обосновываем необходимость использования алгоритмов машинного обучения для прогнозирования осложнений перитонита. В следующем разделе будет описан набор данных, использованный в этом исследовании, а также шаги предварительной обработки, предпринятые для очистки и форматирования данных для анализа.

# Глава 2. Исследование и обработка данных

Наши коллеги в области медицины собрали набор данных, состоящий из 1600 записей, каждая из которых содержит различные показатели и информацию о пациентах. Показатели включали общий анализ мочи, общий анализ крови, биохимия крови, посевы мочи, УЗИ брюшной области, возраст пациента, характер перитонита и другие анализы. Целевой переменной было наличие или отсутствие осложнений.

Изначально мы столкнулись с трудностями из-за человеческого фактора. Нам было отправлено множество дубликатов записей, перепутанных и неверных данных. В конце концов, коллеги предоставили чистый набор данных, не содержащий дубликатов и не релевантной информации.

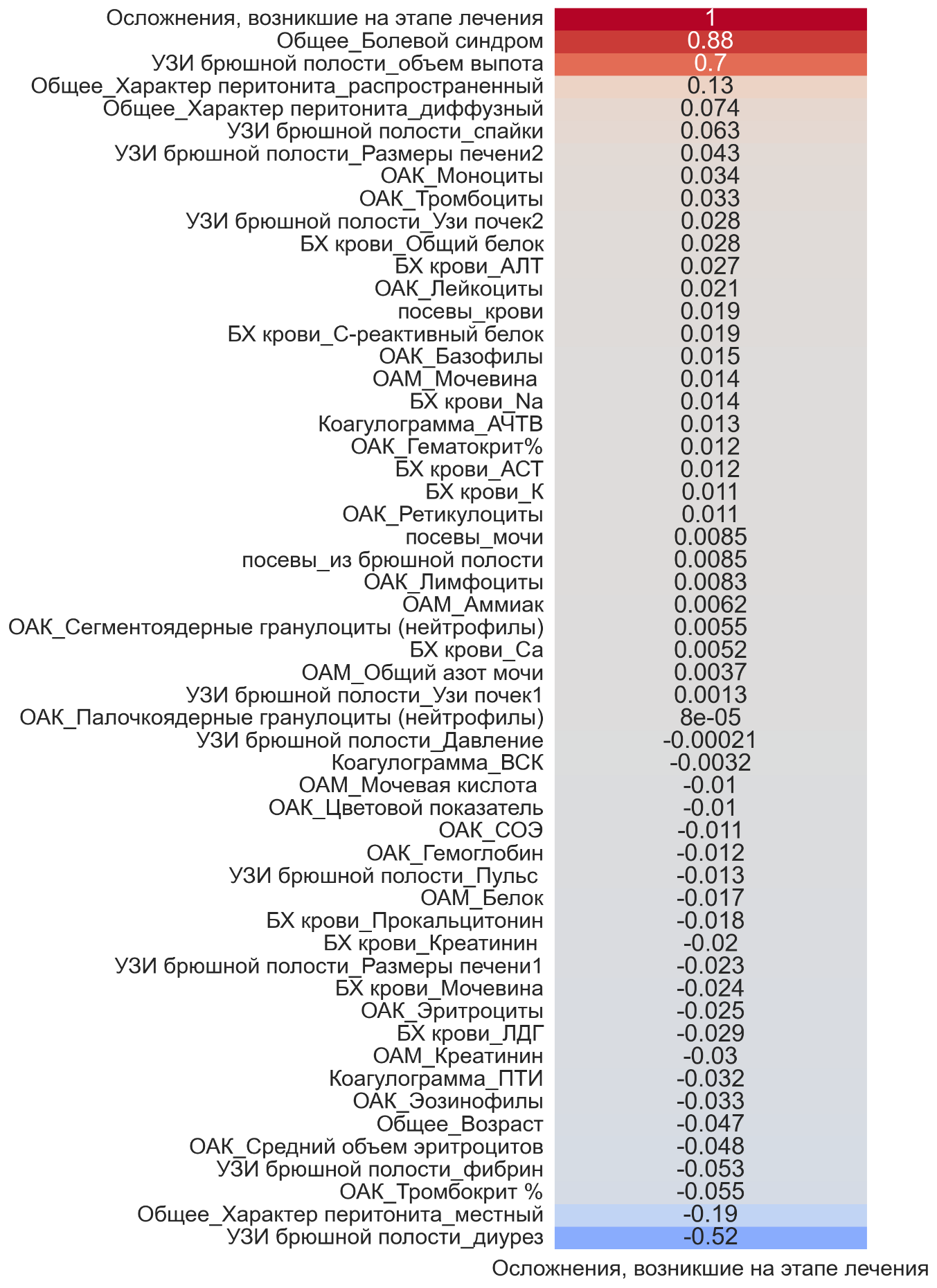
При первоначальном исследовании мы не обнаружили отсутствующих данных или повторяющихся записей. Мы удалили столбцы, лишенные уникальных значений, и удалили столбцы, содержащие постфактум или крайне субъективную информацию. Поскольку существовало несколько известных типов осложнений, мы определили целевую переменную из двух категорий – возникновение или отсутствие каких–либо осложнений, преобразовав исходную проблему с несколькими классами (различные типы осложнений) в проблему бинарной классификации. Анализ распределения числовых и категориальных признаков выявил дисбаланс в целевой переменной: только 35% записей указывают на осложнения.



**Рис. 1**. Распределение целевого признака.

Мы решили не удалять значения выбросов, считая их естественными в наборах медицинских данных. Впоследствии мы закодировали категориальные данные, используя one hot кодирование.

Построение корреляционной матрицы Пирсона позволило нам наблюдать сильные корреляции между целевой переменной и несколькими признаками. Три признака показали очень высокие корреляции, причем болевой синдром был самым сильным - 0,88.



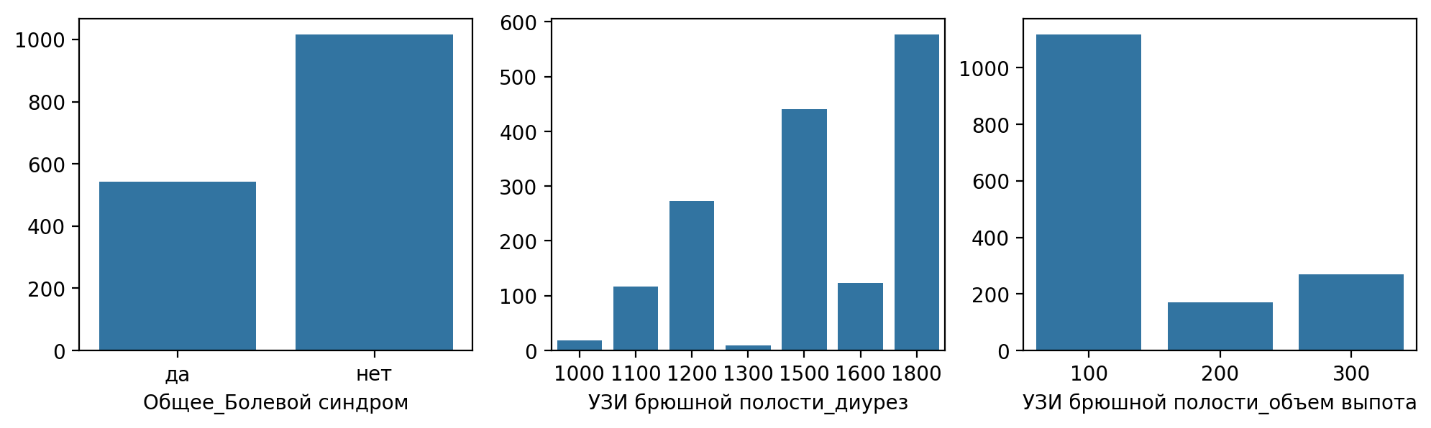
**Рис. 2.** Корреляция с целевым признаком.

Использование mutual information из Sklearn для выявления нелинейных взаимосвязей подтвердило основные характеристики, выявленные корреляционной матрицей Пирсона. Учитывая высокую корреляцию с болевым синдромом, достижение точности в 88% стало возможным просто за счет постоянного прогнозирования осложнений при его наличии. Теперь наша цель сместилась в сторону повышения качества прогнозирования за счет использования оставшейся информации.

# Глава 3. Построение моделей

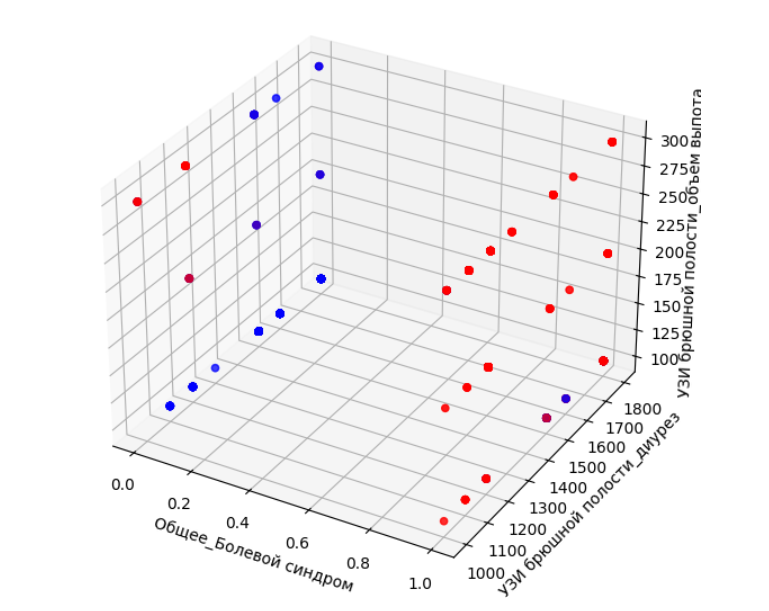
В поисках надежной прогностической модели мы оценили три наиболее непохожих друг на друга и общепризнанно самых эффективных алгоритма классического машинного обучения: K-ближайших соседей (KNN), машины опорных векторов (SVM) и градиентный бустинг на решающий деревьях. После подготовки нашего набора данных, разделения его на обучающие, валидационные и тестовые наборы и выполнения нормализации данных для KNN и SVM мы провели k-кратную перекрестную проверку наряду с рандомизированным поиском для оптимизации гиперпараметров.

В ходе этого процесса мы выявили в наших данных критическую особенность, связанную с тремя наиболее важными признаками: диурезом, объемом выпота и болевым синдромом. Из-за их ограниченных вариаций эти факторы генерировали всего 200 уникальных комбинаций, что создавало проблему, при которой дополнительные данные, если обучать модель и делать предсказание только по этим трем признакам, представляли собой дубликаты. Эта особенность требовала полного изменения подхода к обучению.

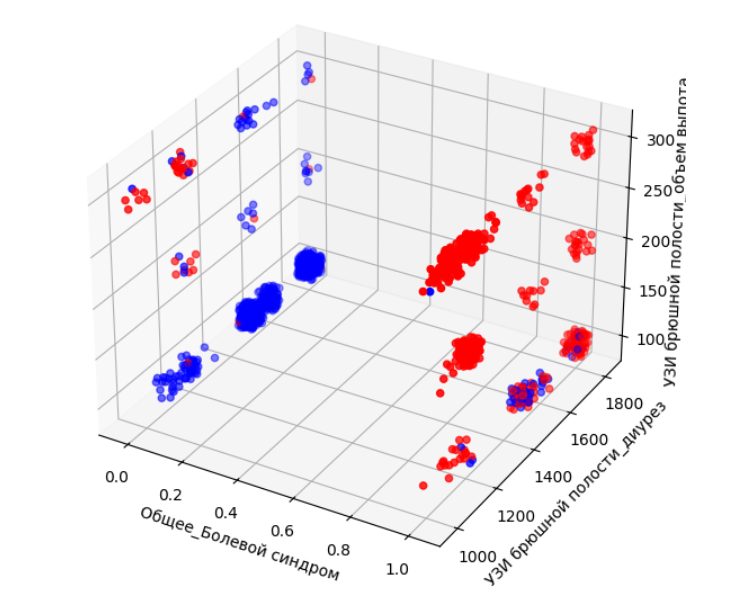


**Рис. 3.** Ограниченные вариации наиболее важных признаков.

Столкнувшись с этим препятствием, мы разработали два решения, показавших свою эффективность. Первый подход заключался в использовании возможностей KNN, поскольку мы исходили из того, что идентичные исходные данные приведут к дублированию записей, которые при этом будут иметь разные метки классов, что делает оценку вероятности возникновения каждого набора из трёх признаков необходимой для определения распространенности. Если, скажем, у 95 человек возникнут осложнения при заданном наборе трёх признаков, а у 5 других, при таком же наборе трёх признаков не возникнут осложнения, то, с точки зрения статистики, правильный ответ будет указывать на осложнения в 95% случаев. По совпадению, KNN вычисляет вероятности точно так, как описано. Просто в данном случае большая часть соседей будет совпадать с входным экземпляром, находясь в пространстве признаков в одной и той же точке, расстояние до которой будет равно 0.

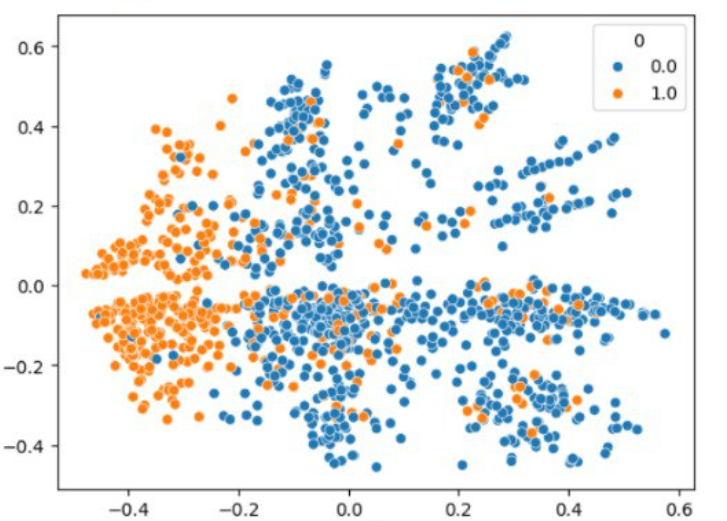


**Рис. 4**. Признаковое пространство с перекрывающимися точками.



**Рис. 5.** Признаковое пространство с принудительно зашумленными точками.

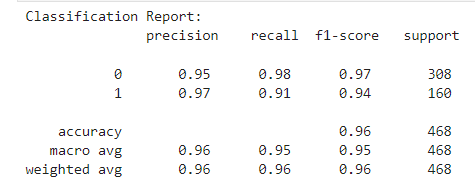
Во втором подходе, мы реализовали своеобразный feature engineering, объединив другие менее значимые признаки с нашими тремя доминирующими, применив далее нелинейный метод главных компонент (Kernel PCA) для получения новых непрерывных признаков, которые уже можно передать бустингу без опасения присутствия дубликатов.



**Рис. 6.** Признаковое пространство после понижения размерности. *Для иллюстрационных целей понижение сделано до двухмерного пространства. Также в данном понижении не использовался самый важный признак – болевой синдром. Он после понижения добавляется в качество нового столбца к получившимся данным*

Оба подхода продемонстрировали схожую эффективность, но был выбран подход с KNN, т.к. он использовал меньше признаков и был более интерпретируемый.

Используя KNN, мы достигли значения ROC-AUC, равного 0,98, при этом показатели precision и recall колебались в пределах ±0,95 для положительного класса, демонстрируя незначительные колебания, обусловленные изменением разбиения набора данных на обучающие и тестовые наборы.



**Рис. 7.** Метрики на тестовом множестве.

В целом, предложенные методы успешно справились с уникальными задачами, связанными с нашим набором данных, что привело к получению точного и надежного прогнозирующего решения.

В заключение, в этой главе проиллюстрирован процесс проб и ошибок, связанный с выбором подходящих алгоритмов машинного обучения, точной настройкой гиперпараметров и учета особенностей, присущих нашему специфическому набору данных. Представленные решения дали обнадеживающие результаты, укрепив нашу уверенность в конечной эффективности выбранного подхода.

# Глава 4. Развертывание модели

Одна из целей данной работы это развертывание модели в виде веб приложения для удобного практического использования. Первоначально для решения поставленной задаче в центре внимания оказались библиотеки Django и Flask. Но позже предпочтение было отдано библиотекам, которые специализировались именно на быстром развертывании моделей машинного обучения, без необходимости в глубоком погружении в frontend и backend. Библиотекой, отвечающей всем поставленным критериям, была Streamlit. Изучение данной библиотеки заняло всего один вечер, благодаря чему уже в скором времени в веб была загружена обученная ранее модель.

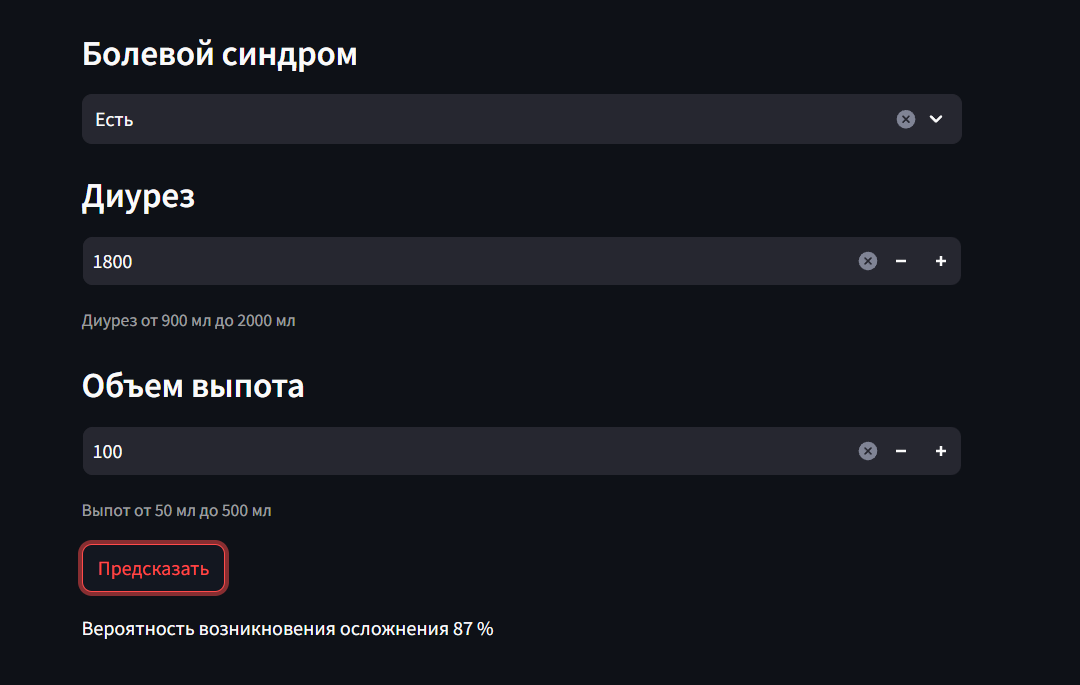
Приложение состоит из серверной части, которая принимает и отправляет запросы, и клиентской части, которая взаимодействует с пользователем.

Серверная часть приложения включает в себя следующие компоненты:

* Приложение Streamlit: это основной интерфейс для взаимодействия с пользователем. Он принимает входные данные от пользователя, передает их модели машинного обучения и возвращает результаты пользователю.
* Модель машинного обучения: это сердце приложения. Модель была обучена на основе большого набора данных и используется для предсказания результатов на основе входных данных, предоставленных пользователем.
* Веса модели: это параметры модели, которые были определены в процессе обучения. Они сохраняются и загружаются каждый раз, когда сервер запускается, чтобы модель могла использовать их для предсказаний.

Клиентская часть — это интерфейс, который пользователи видят и с которым взаимодействуют. Он состоит из формы ввода для предоставления входных данных и области вывода для отображения результатов, возвращаемых сервером.

Также, по согласованию с медиками, были реализованы подсказки и ограничения на некорректный ввод. Далее приложение было поставлено на бесплатный сервер для автономной работы. (Тут, наверно, про докеризацию будем писать)



**Рис. 8**. Интерфейс веб приложения.

# Глава 5. Заключение

В данной работе представлен новый подход к прогнозированию осложнений перитонита с использованием алгоритмов машинного обучения. Была проведена тщательная работа по сбору, обработке и анализу клинических данных о пациентах с перитонитом. Были предприняты шаги по предварительной обработке и подготовке данных к моделированию.

Тремя выбранными алгоритмами машинного обучения для сравнения их производительности в прогнозировании осложнений перитонита были опорные вектора SVM, градиентный бустинг и k-ближайших соседей (KNN). После тщательного анализа и сравнения, наиболее эффективным алгоритмом оказался KNN, который был настроен с использованием перекрестной проверки и поиска гиперпараметров по сетке. Результаты демонстрируют высокую степень точности и надёжности прогнозирующего решения.

Было разработано веб-приложение с использованием библиотеки Streamlit, которое позволяет врачам вводить информацию о пациентах и получать мгновенную обратную связь относительно прогнозируемой вероятности развития осложнений. Это представляет собой значительный шаг на пути улучшения диагностики перитонита и результатов лечения.

В качестве будущей работы можно рассмотреть расширение объёма набора данных, включение дополнительных клинических переменных и изучение альтернативных подходов к машинному обучению. Кроме того, необходимо продолжать мониторинг и обновление модели, учитывая изменяющиеся тенденции в данных и медицинской практике.

В целом, эта работа демонстрирует потенциал использования алгоритмов машинного обучения для улучшения прогнозирования осложнений перитонита, тем самым способствуя повышению качества медицинского обслуживания и лечения пациентов.

# Глава 6. Библиографический список

1. Михайличенко В.Ю., Воронков Д.Е., Кисляков В.В., Цап А.А. Лечение тяжелых форм распространенного гнойного перитонита. *Таврический медико-биологический вестник.* 2022; 25 (1): 20-26.
2. Алипов В.В., Тахмезов А.Э., Полиданов М.А., Мусаелян А.Г., Кондрашкин И.Е., Волков К.А., Алипов А.И. Улучшение результатов лечения и диагностики послеоперационных осложнений в абдоминальной хирургии с применением многофункционального устройства. *Медицинская наука и образование Урала.* 2023; 24 (1-113): 67-71
3. Gandhi K., Prasad D., Malhotra V., Agrawal D., Beniwal P., Leon Dsouza A.V. Clinicoepidemiological profile of peritonitis complicating acute peritoneal dialysis: A single-center experience. *Saudi J Kidney Dis Transpl.* 2017; 28 (5):1133-1137.
4. Савельев В.С., Филимонов М.И., Подачин П.В., Чубченко С.В. Ошибки выбора тактики хирургического лечения распространенного перитонита. *Анналы хирургии*. 2008; 1: 26-32
5. Барулина М.А., Масляков В.В., Полиданов М.А., Романов Р.А., Волков К.А. Математические и алгоритмические методы исследования признаков возникновения осложнений при перитоните*. Математическое моделирование, компьютерный и натурный эксперимент в естественных науках*. 2023; 2. Ссылка активна на 22.03.2024: mathmod.esrae.ru/42-170
6. Полиданов М.А., Волков К.А., Медунов А.С., Масляков В.В., Барулина М.А. Прогнозирование осложнений при перитоните на раннем этапе математическими и алгоритмическими методами. В книге: Инновационные технологии в медицине: взгляд молодого специалиста. Сборник докладов IX Всероссийской научной конференции молодых специалистов, аспирантов, ординаторов. Рязань. 2023; 79-81
7. Bаlаnescu L., Bаetu A.E., Cardoneanu A.M., Moga A.A., Bălănescu R.N. Predictors of Complicated Appendicitis with Evolution to Appendicular Peritonitis in Pediatric Patients. *Medicina (Kaunas)*. 2022; 59 (1): 21.
8. Сараев А.Р. Усовершенствование классификации перитонитов. *Вестник Авиценны.* 2018; 20(4):442-447
9. Клинические рекомендации. Острый перитонит. Утверждены Минздравом РФ. Кодирование по Международной статистической классификации болезней и проблем, связанных со здоровьем: K65.0/K65.8/K65.9/K67/K67.0/K67.1/K67.2/K67.3. Год утверждения: 2023. Ссылка активна на 22.03.2024: <https://clck.ru/39cMnx>
10. Яндекс. Книга по машинному обучению. Ссылка активна на 19.04.2024: https://education.yandex.ru/handbook/ml
11. Pearson correlation criterion. The link is active as of 22.03.2024: https://medstatistic.ru/methods/methods8.html?ysclid=lgce2phu7s101348257
12. CatBoostThe link is active as of 22.03.2024: <https://catboost.ai/?ysclid=lgce8l4u2w627918923>
13. Streamlit. A faster way to build and share data apps. The link is active as of 19.04.2024: <https://streamlit.io/>
14. Scikit-learn. Machine Learning in Python. The link is active as of 19.04.2024: https://scikit-learn.org/stable/index.html
15. Свидетельство о регистрации программы для ЭВМ № 2024612173 / 29.01.2024. Заявка от 25.12.2023. Барулина М.А., Полиданов М.А., Сухой Д.В., Волков К.А., Масляков В.В., Паршин А.В., Высоцкий Л.И. Система предсказания вероятности наступления осложнений после проведения операции больным с перитонитом. Ссылка активна на 22.03.2024: <https://www.elibrary.ru/download/elibrary_60782434_97687051.PDF>
16. Свидетельство о регистрации программы для ЭВМ № 2024615898 / 13.03.2024. Заявка от 03.03.2024. Барулина М.А., Полиданов М.А., Сухой Д.В., Волков К.А., Масляков В.В., Паршин А.В., Высоцкий Л.И. Интернет-сервис для предсказания вероятности наступления в послеоперационных осложнений у больных с перитонитом. Ссылка активна на 22.03.2024: <https://fips.ru/EGD/098da9a5-5ad1-45d3-9376-da9cdb9cd898>