***М.М. Попович, М.В. Григорьев, А.В. Макарихин***

*Тюменский государственный университет, г. Тюмень*

**УДК 519.257**

# РАЗРАБОТКА ПОДСИСТЕМЫ ТРИАЖА ПАЦИЕНТОВ НА ПЛАНОВУЮ ОПЕРАЦИЮ СЕРДЦА С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

**Аннотация.** В работе представлен опыт разработки подсистемы триажа пациентов на плановую операцию сердца в региональной медицинской информационной системе Тюменской области с использованием машинного обучения. Целью данной работы является повышение эффективности взаимодействия медицинского персонала с пациентами в медицинской информационной системе. В работе отражены результаты проектирования подсистемы триажа, составлены подробные схемы каждого модуля. Автор описывает все этапы построения моделей машинного обучения для прогнозирования периода пребывания пациента в стационаре и прогнозирования вероятности выздоровления. В работе отражен опыт использования облачного сервиса IBM Cloud Pak и применения в качестве алгоритма градиентного бустинга.

**Ключевые слова:** машинное обучение, медицина, data science, 1С, IBM Cloud Pak.

## Введение

Автоматизация процессов, протекающих в жизнедеятельности объектов здравоохранения, существенно повышает эффективность их работы, что напрямую влияет на качество медицинского обслуживания пациентов. Оснащение медицинских учреждений качественным и современным программным обеспечением позволяет упростить задачу хранения и обработки большого количества данных, требуемых для качественного обслуживания пациентов. Вследствие этого, продуктивность врачебного персонала повышается

и большинство внутренних процессов протекают без участия человека на основе автоматизации. Находят свое применение и современные технологии интеллектуального анализа данных. Подтверждением этому является внедрение на федеральном уровне вертикально-интегрированной медицинской информационной системы (ВИМИС) с 2020 года. На данный момент, одна из задач ВИМИС в области анализа данных – это прогнозирование онкологических заболеваний, но повсеместно ВИМИС захватывает и другие медицинские области.

В региональной медицинской информационной системе Тюменской области, начиная с 2015 года, накопился достаточный объем данных, нуждающихся в изучении и их анализе. Ранее, уже применялся экспериментальный опыт разработки модели машинного обучения, связанный с оценкой рисков сердечно-сосудистых заболеваний. Точность построенной модели составила 97%. Со временем открываются новые области и идеи для применения интеллектуального анализа данных. Прежде всего, они достигаются путем общения напрямую с пользователями системы – врачебным персоналом. Но следует понимать, что медицина – сложная и важная область в нашей жизни, и применение таких новшеств должно происходить осознано. Использование результатов машинного обучения должно носить рекомендательный характер.

## 1. Постановка задачи

Подсистема триажа пациентов на плановые операции с использованием интеллектуального анализа данных предназначена для медицинской сортировки пациентов, ожидающих проведения оперативных вмешательств, с учетом их первичных показателей здоровья и других значимых факторов. Главная идея подсистемы – создание инструмента для организации электронной очереди на операции с применением машинного обучения.

Целью создания данной подсистемы является повышение эффективности взаимодействия медицинского персонала с пациентами в медицинской информационной системе.

Понятие «эффективность» включает в себя три ключевых показателя:

– уменьшение затрачиваемого времени на регистрацию пациента для проведения оперативных вмешательств;

– повышение пропускной способности отделения;

– сокращение числа промежуточных бюрократических этапов от первичного приема до проведения операции;

Для достижения поставленной цели были определены следующие задачи:

– создание модели для прогнозирования периода пребывания пациента в стационаре;

– создание модели для прогнозирования вероятности выздоровления пациента;

– разработка модуля прогнозирования нагрузки коечного фонда отделения;

– разработки модуля очередей пациентов на оперативные вмешательства.

Перед разработкой подсистемы было проведено интервьюирование медицинского персонала, который будет работать с системой. В ходе обсуждения были выявлены необходимые требования к подсистеме и составлена контекстная диаграмма. Были выявлены четыре роли пользователей, которые будут непосредственно работать с подсистемой – врач, заведующий отделением, главный врач (руководители МО) и сотрудник МИАЦ. Главный врач и МИАЦ получают из системы отчеты в срезе МО, отделений, номенклатур медицинских услуг (НМУ) и по полисам (ОМС, ДМС и т.д.), также главный врач может просматривать прогнозы нагрузки на отделения. Добавлением пациентов и изменением их номеров в очереди занимаются врач и заведующий отделением. Заведующий имеет дополнительную возможность исключать пациентов из очереди и прогнозировать нагрузку на отделение.

## 2. Проектирование подсистемы триажа пациентов на операции

Подсистема триажа пациентов на плановые операции – это подсистема, предназначенная для организации электронной очереди и прогнозирования состояния коечного фонда по отделениям с применением методов интеллектуального анализа данных. Данная подсистема является модульной, работа которых может не зависеть друг от друга.

Обработка и хранение полученных данных из моделей машинного обучения производится внутри медицинской информационной системы (МИС) на базе платформы 1С.

Функционирование моделей ML находится вне контура МИС.

На рисунке 1 изображена общая схема разрабатываемой подсистемы триажа пациентов.

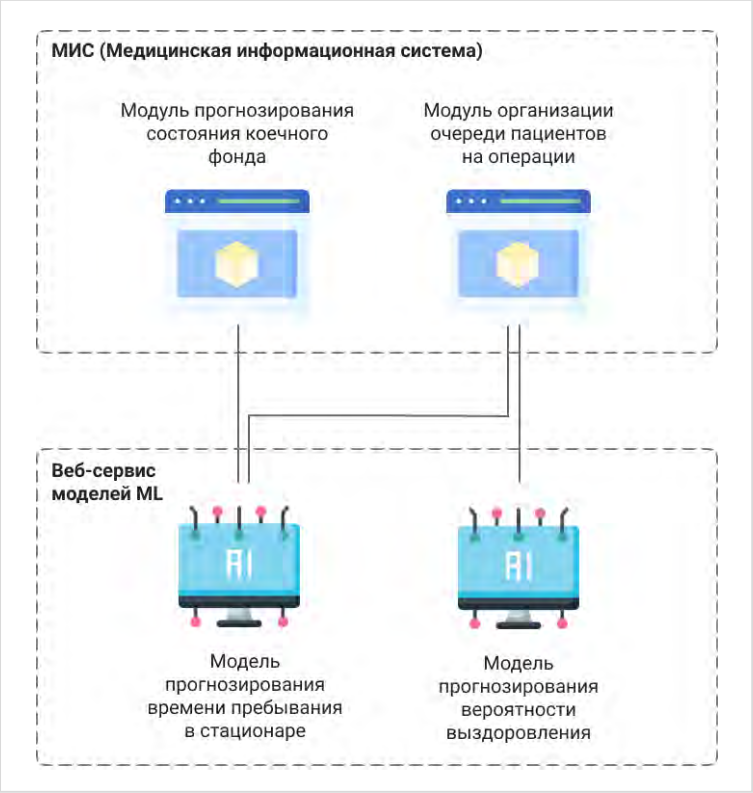


*Рис. 1.* Общая схема структуры подсистемы триажа пациентов

## 3. Модуль прогнозирования нагрузки коечного фонда отделения

Данный модуль предназначен для прогнозирования загруженности отделения по количеству занятых и свободных коечных мест. При прогнозировании собирается первичная информация по каждому пациенту,

который находится на лечении в стационаре и по результатам, формируется возможное состояние коечного фонда на указанную дату прогноза. Пациенты в отделение могут поступать несколькими способами: планово, экстренно и путем перевода из других отделений. Для обеспечения запаса мест для экстренных пациентов, модуль учитывает процент резервных коек по каждому из отделений. Наглядная схема работы изображена на рисунке 2.

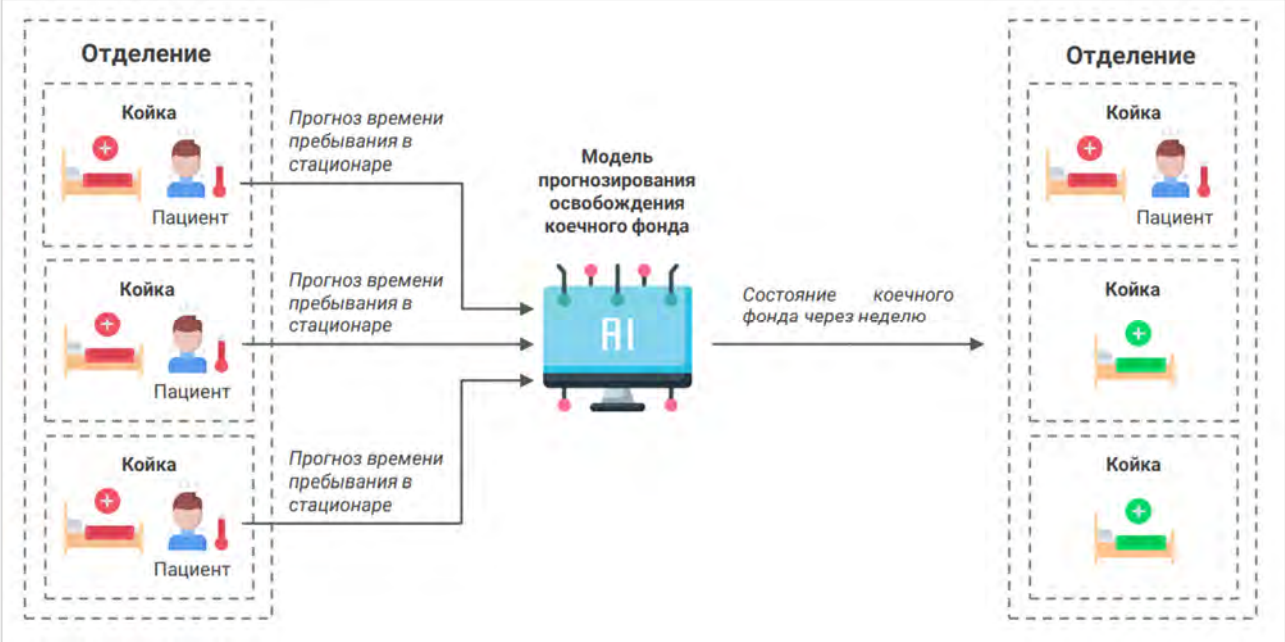


*Рис. 2.* Схема работы модуля прогнозирования состояния коечного фонда

### 4. Модуль организации очереди пациентов на операции

Работа данного модуля направлена на обеспечение локального инструмента МИС, предназначенного для организации очереди пациентов на различные операции. Ответственное лицо имеет возможность использовать модуль как вручную, так и в автоматическом режиме, с помощью оценки «рейтинга» пациента обученной моделью машинного обучения. При добавлении пациента в очереди, по нему собирается первичная информация, необходимая для прогноза вероятности его выздоровления и его периода пребывания в стационаре. Затем, данные передаются в модели, которые возвращает результат в виде вероятности выздоровления пациента и периода его пребывания в стационаре. По окончанию, результирующая функция вычисляет «рейтинг» пациента и назначает ему номер в очереди. Ответственное лицо, или врач, может

скорректировать итоговый результат, и добавить пациента в очередь вручную. Общая схема работы модуля представлена на рисунке 3.

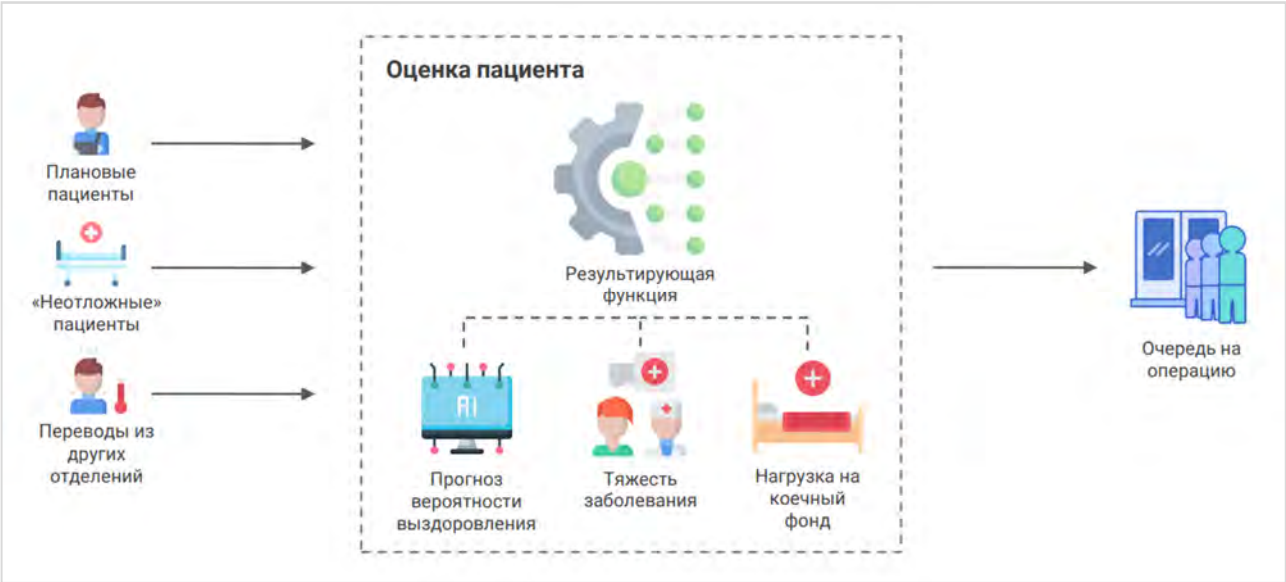


*Рис. 3.* Схема работы модуля организации очереди пациентов на операцию

## 5. Построение модели машинного обучения «Прогнозирование периода пребывания пациента в стационаре»

Поскольку основной задачей данной модели является прогнозирование периода пребывания пациента в стационаре, то был собран ряд данных, связанных непосредственно с его госпитализацией. Сбор данных осуществлялся в разрезе случаев (медицинских карт). Данные выгружались с рабочей базы Областной клинической больницы №1 города Тюмени по кардиохирургическому отделению в период с 1 января 2018 года. Данные пациентов были деперсонифицированы путем выгрузки их системных идентификаторов. В МИС, на данный момент, записывается большой объем данных по пациенту в период его лечения в стационаре, но для решения данной задачи было принято решение выделить следующий ряд данных, загружаемых из МИС: пол, дата рождения, дата госпитализации, диагнозы, дата выписки. Дальнейшая обработка данных для обучения модели осуществлялась в среде Python, с использованием библиотек: Pandas и LabelEncoder.

Признак «Пол» был преобразован в категориальный признак «SEX» по принципу: «0» – женский пол, «1» – мужской пол. Большинство записей (около 67%) принадлежит мужскому полу.

Признак «Дата рождения» преобразован в непрерывный признак «AGE» (возраст). Большая часть пациентов относится к возрастной группе от 50 до 70 лет.

Признак «Дата госпитализации» был преобразован в категориальный параметр «MONTH_OF_HOSPITALIZATION» (месяц госпитализации), закодированный по порядковому номер месяца в году. Все значения распределены равномерно, но в наиболее популярными являются летние и осенние месяцы.

Признак «Дата выписки» изначально был преобразован в непрерывный признак «Время пребывания в стационаре» путем нахождения разности между двумя датами: госпитализации и выписки, а затем был сформирован новый признак «TIME_OF_HOSPITALIZATION_INTERVAL», который разделен на 3 класса: «до 5 дней», «от 5 до 14 дней», «от 14 дней». Разделение на такие классы было согласовано с заказчиком. Этот признак, в дальнейшем будет являться целевым.

В МИС диагнозы хранятся в международном формате МКБ-10. Например, диагноз с кодом «I21.3», подразумевает под собой острый трансмуральный инфаркт миокарда неуточненной локализации. При обсуждении возможных вариантов преобразования этого параметра с врачебным персоналом, было принято решение подсчитывать количество диагнозов по родительскому диагнозу в рамках одного случая. Т.е. был взят только родительский диагноз до точки, в результате из кода «I21.3» был получен диагноз «I21». Количество уникальных значений диагноза составило – 474 родительских кодов МКБ-10.

Далее, с помощью Dummy-кодирования, для данного признака были созданы N новых признаков по шаблону «DIAGNOSIS_%диагноз%», где N – число уникальных диагнозов из выборки и %диагноз% – один из уникальных диагнозов.

По окончанию предобработки, было получено 5430 записей. Количество признаков – 495.

Для экспериментального моделирования использовалась платформа IBM Cloud Pak for Data.

В качестве тестируемых алгоритмов были выбраны классификаторы дерево решений, случайный лес и градиентный бустинг. Наиболее оптимальные показатели показал последний алгоритм – градиентный бустинг, поэтому было принято решение выбрать его в качестве основного и подбирать гиперпараметры относительно него.

Для повышения точности модели важную роль сыграл параметр «Максимальная глубина» (max_depth), поскольку в исходном наборе данных большое количество признаков. Итоговые гиперпараметры модели получились следующие: число деревьев (n_estimators) – 100, Размер шага (eta) – 0.3, Минимальное изменение значения loss функции для разделения листа на поддеревья (gamma) – 0, максимальная глубина дерева (max_depth) – 30, L2 регуляризация (lambda) – 1, L1 регуляризация (alpha) – 1.

В результате, модель показала хорошую точность, процент правильно определенных классов на обучающей выборке составил 85.43%, на тестовой – 84.3%. Соотношение обучающей и тестовой выборки – 90 и 10% соответственно.

## 6. Построение модели машинного обучения «Прогнозирование вероятности выздоровления»

Прогнозирование вероятности выздоровления пациента должно иметь высокую точность, поскольку от этого зависит план лечения пациента. На данный момент, в МИС содержится большой объем неструктурированных медицинских документов, данные которых вводятся врачами в свободной форме. С недавнего времени происходит внедрение вертикально-интегрированной медицинской системы на федеральном уровне, которая предназначена для сбора структурированных сведений по пациентам. Поэтому, сейчас правильно подготовленных данных не так много, и в рамках данной

работы было принято решение, провести эксперименты с имеющимися параметризированными данными – показателями здоровья пациента по общему анализу крови (ОАК).

Задача данной модели заключалась в бинарной классификации пациентов по признаку «Выздоровление» (да/нет). В качестве предиктора выздоровления, по результатам обсуждения с врачебным персоналом, было принято решение использовать значение фракции выброса левого желудочка (ФВЛЖ). Данный параметр эхокардиографии отвечает за процент выброса крови в сосуды из левого желудочка сердца.

Так как главная задача модели состоит в бинарной классификации, то было принято решение преобразовать данный показатель в бинарный признак «Выздоровление» (1 – да, 0 – нет). К выздоровлению относятся те случаи, в которых ФВЛЖ перед выпиской находится в норме (> 55%).

Для подготовки данных для обучения модели была произведена выгрузка с рабочей базы ОКБ-1 за период с 1 января 2018 года по всем пациентам, у которых имелся медицинский документ «Протокол эхокардиографии» и присутствуют показатели здоровья по ОАК и показания диастолического/систолическое давления перед выпиской.

Были выгружены следующие признаки: пол, дата рождения, вид транспортировки, состояние, срочность госпитализации, диагнозы, диастолическое давление, систолическое давление, процент эозинофилов, гематокрит, процент базофилов, процент моноцитов, процент лимфоцитов, процент нейтрофилов, базофилы, эозинофилы, лимфоциты, моноциты, нейтрофилы, тромбокрит, средний объем тромб, относительная ширина распределения тромбоцитов, ширина распределения эритроцитов, тромбоциты.

Признак «Пол» был также как и в первой модели преобразован в категориальный признак «SEX» по принципу: «0» – женский пол, «1» – мужской пол.

Аналогично, признак «Дата рождения» преобразован в непрерывный признак «AGE» (возраст).

Признаки «Вид транспортировки», «Состояние», «Срочность госпитализации» преобразованы в категориальные. Большинство пациентов поступало в плановом порядке, в удовлетворительном состоянии и способными передвигаться самостоятельно.

Показатели здоровья не изменялись и не преобразовывались.

Диагнозы были обработаны тем же способом, что и в первой модели (см. п. 5).

ФВЛЖ, как уже было указано ранее, преобразован в бинарный классовый признак «Выздоровление» (1 – да, 0 – нет). В небольшой степени преобладают выздоровевшие пациенты.

В результате было выгружено 1023 записи с 382 признаками.

Моделирование производилось так же, как и с первой моделью, с помощью сервиса IBM Cloud Pak for Data.

В результате, модель показала хорошую точность, процент правильно определенных классов на обучающей выборке составил 90,74%, на тестовой – 91,63%. Соотношение обучающей и тестовой выборки – 90 и 10% соответственно. Итоговые гиперпараметры модели получились следующие: число деревьев (n_estimators) – 100, Размер шага (eta) – 0.3, Минимальное изменение значения loss функции для разделения листа на поддеревья (gamma) – 0, максимальная глубина дерева (max_depth) – 15, L2 регуляризация (lambda) – 1, L1 регуляризация (alpha) – 0.

**Заключение**

Процесс реализации данного проекта включает два важных аспекта: проектирование и реализация.

Первый аспект – проектирование подсистемы включал в себя анализ и разработку различных спецификаций к проектируемой системе. На данном этапе должны быть сформулированы перечни специфических особенностей проектируемой системы, а также разработаны диаграммы, отражающие ключевые моменты в процессе жизнедеятельности системы. Также были

подобраны технологии и обсуждены цели и задачи каждого компонента системы.

Второй аспект, подразумевал под собой создание моделей машинного обучения, разработку пользовательского интерфейса и программную реализацию всех компонентов. Для построения моделей машинного обучения был произведен сбор данных с рабочих баз ОКБ-1, на основе которых происходило обучение запланированных моделей. Полученные модели были размещены на выделенном виртуальном сервере, вне контура МИС. Далее, на основе на предыдущего этапа проектирования разрабатывалась подсистема, обладающая, необходимой в рамках поставленных задач, функциональностью.

За дальнейшим развитием системы стоит тестирование, с помощью которого будут дорабатываться упущенные моменты на стадиях проектирования и разработки. Заключительными этапами разработки являются внедрение и эксплуатация продукта. Поскольку данная система является частью единой большой системы, от которой зависит множество факторов, влияющих на процессы внутри медицинских учреждений, то вопрос об интеграции требует тщательного анализа и проверок со стороны квалифицированных экспертов.

## СПИСОК ЛИТЕРАТУРЫ

1. Асатрян А., Голиков А. Методическое пособие по эксплуатации крупных информационных систем на платформе «1С:Предприятие 8». 2-е изд. М.: 1С-Паблишинг, 2017. 331 с.

2. Вендров А. CASE-технологии. Современные методы и средства проектирования информационных систем. М.: Финансы и статистика, 2005. 1410 с.

3. Международная статистическая классификация болезней и проблем, связанных со здоровьем (10-й пересмотр) // Министерство здравоохранения Российской Федерации (НСИ). URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1005/version/2.18 (дата обращения: 21.05.2021).

4. IBM Cloud Pak for Data // IBM. URL: https://www.ibm.com/ru-ru/products/cloud-pak-for-data (дата обращения: 20.03.2021).

5. XGBoost Documentation // Read the Docs. URL: https://xgboost. readthedocs.io/ en/release_0.90/ (дата обращения: 23.04.2021).

6. Обзор системы «1С:Предприятие 8» // 1С. URL: https://v8.1c.ru/tekhnologii/overview/ (дата обращения: 01.06.2021).