УДК 314.48 DOI: 10.26102/2310-6018/2020.31.4.025

Применение методов машинного обучения при назначении терапии гипертонической болезни

М.А. Фирюлина[1], И.Л. Каширина[1], Е.Я. Гафанович[2]
[1]Воронежский государственный университет, Воронеж, Россия
[2]Саратовский государственный медицинский университет им. В.И. Разумовского, Саратов, Российская Федерация

Резюме: Несмотря появление новых современных препаратов, показатели смертности от гипертонической болезни остаются высокими. Эта проблема, в частности, обусловлена тем, что для эффективного лечения этого заболевания необходима комбинация нескольких групп препаратов. Целью данного исследования является разработка моделей автоматизированного подбора препаратов для лечения гипертонии на основе индивидуальных характеристик пациента, а также оценки эффективности назначенного лечения на основе имеющихся клинических показателей пациентов и предполагаемой комбинации препаратов. Исходная выборка данных содержит деперсонифицированную информацию о 262 пациентах кардиологического стационара по 66 клиническим показателям. Было рассмотрено 6 групп препаратов: БАБ, И-АПФ\АРА, БКК группы нифедипина, БКК группы верапамила, диуретики, препараты центрального действия. Методы машинного обучения использовались для выявления детерминант, способствующих успеху медикаментозного лечения гипертонии для данной выборки пациентов. В ходе исследования для достижения поставленной цели было построено несколько моделей машинного обучения для решения задач классификации и регрессии. Наибольшую точность показали модели градиентного бустинга XGBOOST – для задачи классификации и CATBOOST – для задачи регрессии. По результатам исследования можно сделать выводы, какие клинические показатели наиболее значимы для эффективного лечения каждым из рассматриваемых препаратов.

Ключевые слова: машинное обучение, градиентный бустинг, деревья решений, случайный лес, артериальная гипертония, артериальное давление.

Для цитирования: Фирюлина М.А., Каширина И.Л., Гафанович Е.Я. Применение методов машинного обучения при назначении терапии гипертонической болезни. *Моделирование, оптимизация и информационные технологии.* 2020;8(4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=871 DOI: 10.26102/2310-6018/2020.31.4.025

Using of machine learning methods in prescribing hypertension therapy

M.A. Firyulina[1], I.L. Kashirina[1], E.Y. Gafanovich[2]
[1]Voronezh State University, Voronezh, Russia
[2]Saratovsky State Medical University named after V.I. Razumovsky, Saratov, Russian Federation

Abstract: Despite the emergence of new modern medicines, mortality rates from essential hypertension remain high. This problem is since a combination of several groups of medicines is required to effectively treat this disease. The aim of this study is to develop models for the automated selection of medicines for the treatment of hypertension based on the individual characteristics of the patient, as well as to assess the effectiveness of the prescribed treatment based on the available clinical indicators of patients and the proposed combination of drugs. The original dataset contains depersonalized information on 262 patients of the cardiological hospital for 66 clinical parameters. Six groups of drugs

were considered: BAB, I-ACE\ARA, CCB of the nifedipine group, CCC of the verapamil group, diuretics, centrally acting medicines. Machine learning techniques have been used to identify determinants that contribute to the success of drug treatment for hypertension in each sample of patients. During the study, to achieve this goal, several machine learning models were built to solve classification and regression problems. The highest accuracy was shown by the gradient boosting models XGBOOST for the classification problem and CATBOOST for the regression problem. Based on the results of the study, it can be concluded which clinical indicators are most significant for effective treatment with each of the medicines under consideration.

Keywords: machine learning, gradient boosting, decision trees, random forest, arterial hypertension, arterial pressure.

For citation: Firyulina M.A., Kashirina I.L., Gafanovich E.Y. Using of machine learning methods in prescribing hypertension therapy. *Modeling, optimization and information technology*. 2020;8(4). Available from: https://moitvivt.ru/ru/journal/pdf?id=871 DOI: 10.26102/2310-6018/2020.31.4.025 (In Russ).

1 Введение

Гипертоническая болезнь - распространенное заболевание, которым страдает примерно 20% населения мира [1]. Это одна из основных причин смертности, наряду с инсультом, сердечной недостаточностью, ишемической болезнью сердца и инфарктом миокарда. В дополнение к изменению образа жизни, лечение включает назначение одного или нескольких классов лекарств. К наиболее часто используемым препаратам относятся бета-адреноблокаторы (БАБ), ингибиторы ангиотензинпревращающего фермента (И-АПФ\АРА), блокаторы кальциевых каналов (БКК) группы нифедипина, БКК группы верапамила, диуретики, препараты центрального действия

Группы экспертов-медиков со всего мира регулярно анализируют накопленные данные об успехе терапии артериальной гипертонии [2]. Доказано, что большинству пациентов требуются препараты более чем одного класса, но рекомендации в руководствах по лечению гипертонии неоднородны. Существующие рекомендации основаны, главным образом, на рандомизированных клинических исследованиях (РКИ). Однако ни одно РКИ не выявило оптимальной стратегии дозирования или комбинации препаратов.

Большинство традиционных алгоритмов в медицине - это наборы правил, основанные на экспертных знаниях по определенной теме, в данном случае в области кардиологии. Алгоритмы машинного обучения являются относительно новой областью исследований, направленной на выявление новых и достоверных закономерностей в данных. Машинное обучение включает в себя различные инструменты моделирования, которые предназначены для выявления “скрытых зависимостей” путем изучения тенденций, присутствующих в наборах данных. Цель данного исследования заключалась в том, чтобы проверить возможность использования методов машинного обучения для получения представления о лечении гипертонии.

В ходе работы были построены модели назначения отдельных препаратов, модели прогнозирования эффективности лечения для предполагаемой комбинации препаратов, назначенных пациенту, проанализировано влияние комбинации назначенных препаратов на величину систолического (САД) и диастолического артериального давления (ДАД), определены стратегии применения методов машинного обучения для успешного исхода лечения пациентов.

Модели машинного обучения и визуальный анализ результатов проводился на языке программирования Python с использованием сервиса Google Colab.

2 Материалы и методы

2.1 Описание исходных данных

Для анализа использовалась выборка пациентов с деперсонифицированными данными, поступивших в кардиологический стационар с диагнозом артериальная гипертензия. Всего в исследовании было рассмотрено 262 случая.

На основе показателей САД и ДАД при поступлении и при выписке, было вычислено процентное значение снижения артериального давления. Определена эффективность для лечения САД и ДАД по правилу: если САД снизился на 25% или не более 130 единиц, то лечение эффективно. Если ДАД снизился на 20%, или не более 90 единиц, то лечение эффективно. В конечном итоге определена эффективность лечения в целом: если при выписке лечение было эффективно для САД и ДАД. Лечение эффективно при выписке оказалось для 242 пациентов, не эффективно – 20. Средний возраст мужчин в выборке составил 55 лет, женщин 62 года.

Исходный файл содержал информацию по 72 показателям, представленным в Таблице 1.

Таблица 1 – Исходные показатели пациентов
Table 1 – Patient baseline values

Категориальные признаки	Пол, наличие ожирения, степень АГ (артериальной гипертензии), частота гипертонических кризов, наличие ИМ (инфаркта миокарда) в анамнезе, наличие стенокардии напряжения, наличие кардиалгии на фоне гипертонии, наличие застойных явлений, ХСН (хроническая сердечная недостаточность), наличие и вид мерцательной аритмии, наличие ОНМК (острые нарушения мозгового кровообращения) в анамнезе, наличие ЧМТ (черепно-мозговой травмы), наличие аллергических реакций, ХОБЛ (хроническая обструктивная болезнь легких), сахарный диабет в анамнезе, наличие осложнений сахарного диабета, наличие операций, потребовавших общий наркоз, наследственная отягощенность по АГ, менопауза, тип телосложения, ЭОС (электрическая ось сердца), НЖЭС (наджелудочковая экстрасистолия) на ЭКГ, ЖЭС (желудочковая экстрасистолия) на ЭКГ, НЖТ (наджелудочковая тахикардия), АВ-блокада, СА-блокада, ПБПНПГ, ПБЛНПГ, НБПНПГ, НБЛНПГ (показатели сердечного ритма), лечение до поступления И-АПФ\АРА (ингибиторами ангиотензин превращающего фермента), лечение до поступления БКК (блокаторами кальциевых каналов), лечение до поступления диуретиками, лечения до поступления гипотензивными препаратами центр. действия, лечение до поступления БАБ (бета-адреноблокаторами), непереносимость препаратов, прием в стационаре И-АПФ\АРА, прием в стационаре БАБ, прием в стационаре БКК гр. нифедипина, прием в стационаре БКК гр. верапамила, прием в стационаре диуретиков, прием в стационаре препаратов центрального действия.
Непрерывные признаки	Возраст, ИМТ (индекс массы тела), длительность АГ (артериальной гипертензии), ЧСС (частота сердечных сокращений), PQ, QRS, QT (показатели электрокардиограммы), анемия, лейкоцитопения, тромбоцитопения, СОЭ (скорость оседания эритроцитов), удельный вес мочи, глюкоза, холестерин, креатинин, КДР (конечно-диастолический размер), ФВ (фракция выброса), ЛП (размер левого предсердия), ПП (размер правого предсердия), ПЖ (размер правого желудочка), аорта, ДЛП (дислипидемия), ТМЖП (толщина межжелудочковой перегородки), ТЗСЛЖ (толщина задней стенки левого желудочка), САД при поступлении, ДАД при поступлении, САД при выписке, ДАД при выписке, САД на третий день, ДАД на третий день

В Таблице 2 показано распределение назначаемых препаратов для пациентов.

Таблица 2 – Исследуемые препараты
Table 2 – Investigational medicines

Название препарата	Кол-во пациентов, которым назначен препарат	Кол-во пациентов, которым не назначен препарат
БАБ	151	111
И-АПФ\АРА	178	84
БКК гр. нифедипина	60	202
БКК гр. верапамила	40	222
Диуретики	169	93
Препараты центр. действия	32	230

Исходные данные были проанализированы на наличие пропущенных значений, уникальных значений и коррелирующих показателей. Наличие показателей с такими данными влияют на результаты анализа, снижают скорость обучения модели, интерпретируемость и, главное, способность к обобщению [3]. Пропущенные значения были заменены на медианные, так как удаление записей с пропущенными значением сократит объем выборки.

Наличие сильно коррелирующих показателей приводит к снижению производительности из-за высокой дисперсии и меньшей интерпретируемости модели [4]. Было задано пороговое значение, равное 0.7. Если корреляция между признаками была выше порогового значения, один из признаков исключался. В исходной выборке найдено 4 показателя, которые были исключены из дальнейшего анализа: ожирение, стенокардия напряжения, глюкоза, ТЗСЛЖ. На Рисунке 1 представлен фрагмент матрицы корреляции.

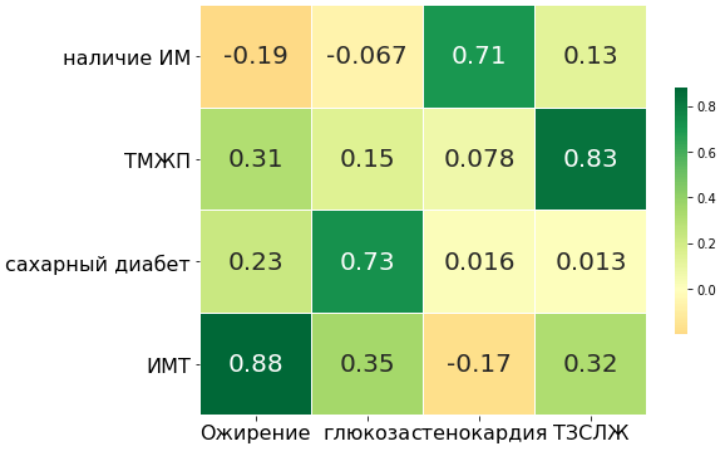

Рисунок 1 – Фрагмент матрицы корреляции
Figure 1 – Fragment of the correlation matrix

2.2 Методы машинного обучения

Для достижения поставленных целей были построены модели машинного обучения для задач классификации и регрессии. Для прогнозирования эффективности лечения при выписке использовалось несколько моделей машинного обучения. Для сравнения были построены модели случайного леса и градиентного бустинга (XGBOOST). Также были построены модели прогнозирования показателей САД и ДАД при выписке и на третий день. Для задачи регрессии наилучшая модель определялась из

следующих: случайный лес, градиентный бустинг (XGBOOST), градиентный бустинг (LIGHTGBM), градиентный бустинг (CATBOOST).

Случайный лес – алгоритм машинного обучения, заключающийся в использовании ансамбля решающих деревьев [5]. Данный алгоритм использует механизм бутстрепа, позволяющий на основе исходного обучающего набора данных с использованием случайного отбора с повторениями сформировать несколько выборок такого же размера. Алгоритм случайного леса можно использовать как для задач регрессии, так и для задачи классификации. При построении регрессионной модели окончательная прогнозируемая величина является средним значением среди всех выходов построенных деревьев. В случае классификации, каждое дерево ансамбля относит классифицируемый объект к одному из классов, и определяется класс, за который проголосовало наибольшее число деревьев. Основным преимуществом данного алгоритма является высокая точность [6].

Градиентный бустинг – это техника машинного обучения, основная идея которой заключается в итеративном процессе последовательного построения частных моделей. Каждая новая модель обучается с использованием информации об ошибках, сделанных на предыдущем этапе, а результирующая функция представляет собой линейную комбинацию всего ансамбля моделей с учетом минимизации некоторой штрафной функции [7]. Данный алгоритм выделяется высокой точностью, в большинстве случаев превосходящей точность остальных методов. Также этот метод устойчив к выбросам, которые часто встречаются в выборках с реальными наблюдениями.

В данном исследовании рассматривались три модели градиентного бустинга: XGBoost, LightGBM и CatBoost. LightGBM и CatBoost довольно новые методы, которые созданы на основе метода градиентного бустинга в попытке улучшения его производительности. Основное различие в том, что в этих моделях используется новая техника односторонней выборки на основе градиента (GOSS) для фильтрации экземпляров данных для нахождения разделения признаков, в то время как XGBoost использует предварительно отсортированный алгоритм и алгоритм на основе гистограмм для вычисления наилучшего разделения [8]. Также в отличии от XGBoost, LightGBM и CatBoost имеют возможность обрабатывать категориальные переменные, CatBoost за счет кодирования категориальных переменных, а LightGBM использует специальный алгоритм, чтобы найти значение разделения категориальных признаков.

2.3 Метрики качества полученных моделей

Так как размер исходной выборки достаточно небольшой, для проверки точности моделей использовалась техника 10-кратной кросс-проверки: исходная выборка разбивалась на 10 частей (фолдов), после чего поочередно одна из этих частей использовалась в качестве тестовой выборки, а остальные 9 – обучающей. Таким образом, процесс обучения модели повторялся десять раз, после чего значения полученных ошибок по всем десяти тестовым выборкам усреднились. При оценке качества моделей для задач классификации использовались метрики, вычисленные на основе матрицы ошибок классификации. На основе элементов матрицы для каждого класса были рассчитаны показатели: (TP) - True Positive (число верно предсказанных примеров класса 1), (FN) - False Negative (число ложноотрицательных примеров, неверно предсказанный класс 0), (TN) - True Negative (верно предсказанный класс 0), (FP) - False Positive (число ложноположительных примеров, неверно предсказанный класс 1). Основная метрика задач классификации – доля правильных ответов, которая вычисляется по формуле 1. Помимо этого были рассмотрены метрики: чувствительность (Sensitivity) – доля истиноположительных примеров от общего числа положительно

предсказанных примеров, и специфичность (Specificity) – доля правильно классифицированных объектов негативного класса. В Формулах 2 и 3 представлен способ расчета этих метрик.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (2)$$

$$Specificity = \frac{TN}{(TN + FP)} \quad (3)$$

Чтобы создать оптимальную диагностическую систему, необходимо найти компромисс между полученными показателями чувствительности и специфичности моделей. Распространенный способ визуализации отношения между этими метриками заключается в использовании ROC-кривой. Величина AUC ROC – площадь под ROC - кривой является компромиссной метрикой, широко применяемой в медицинских исследованиях.

Для оценки точности моделей задач регрессии использовался показатель R^2. Данная величина показывает, насколько условная дисперсия модели отличается от дисперсии реальных значений. Если этот коэффициент близок к 1, то условная дисперсия модели достаточно мала и весьма вероятно, что модель неплохо описывает данные. Если же коэффициент R-квадрат сильно меньше, например, меньше 0.5, то, с большой долей уверенности модель не отражает реальное положение вещей.

Еще одной метрикой в задачах регрессии была RMSE –величина евклидова расстояния между двумя точками, прогнозируемой и исходной. Данный показатель интерпретируется как средняя ошибка модели, на сколько единиц реальное значение в среднем отличается от прогнозируемого. Данный показатель определяется по Формуле 4.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}\ , \quad (4)$$

где N – количество объектов в наблюдении, y_i – фактически ожидаемый результат, $\hat{y}_i$ – прогноз модели.

И третьей вычисляемой метрикой в задачах регрессии была MAE, которая рассчитывается как среднее абсолютных разностей между целевыми значениями и прогнозами. MAE - это линейная оценка, которая означает, что все индивидуальные различия взвешены одинаково в среднем. Преимущество данной метрики в том, что она не так чувствительна к выбросам, как RMSE и R^2. Математически данная метрика рассчитывается по Формуле 5.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|\ , \quad (5)$$

где N – количество объектов в наблюдении, y_i – фактически ожидаемый результат, $\hat{y}_i$ – прогноз модели.

3 Результаты и их обсуждение

3.1 Прогнозирование эффективности лечения заданной комбинацией препаратов

Построение моделей машинного обучения и анализ данных проводилось с помощью библиотек языка программирования Python. Первоначально решалась задача по прогнозированию эффективности лечения при заданном наборе клинических показателей пациента и предполагаемой комбинации препаратов. На выходе прогнозировалась переменная «эффективность при выписке», которая была получена по

правилам, описанным выше. Для решения задачи сравнивалась точность четырех методов машинного обучения: метода случайного леса и трех методов градиентного бустинга (XGBoost, CatBoost, LGBM). В Таблице 3 приведены результаты точности полученных моделей по результатам кросс-валидации.

Таблица 3 – Точность моделей задачи классификации
Table 3 – Accuracy of classification models

Модель МО	*Чувствительн.*	*Специфичность*	*AUC_ROC*	*Accuracy*
XGBoost	1	0.4	0.757	0.95
CatBoost	1	0.3	0.813	0.94
LGBM	0.99	0.4	0.756	0.94
Случайный лес	0.99	0.2	0.77	0.93

Более наглядно соотношение полученных метрик качества для построенных показано на Рисунке 2.

Сравнивая показатели можно сделать вывод, что наиболее точные результаты классификации показали модели XGBoost и LGBM (их графики почти совпали). Такие низкие значения показателя специфичности обусловлены сильной несбалансированностью выборки (всего 20 примеров неэффективного лечения против 242 примеров эффективного). На Рисунке 3 изображены графики ROC-кривых для построенных моделей.

В Таблице 4 показаны по четыре наиболее значимых предиктора для каждой модели при прогнозировании эффективности лечения при выписке. Отметим, что для всех методов значимым показателем, влияющим на эффективность лечения оказался диаметр аорты, для трех из четырех – величина САД при поступлении и величина интервала PQ на электрокардиаграмме.

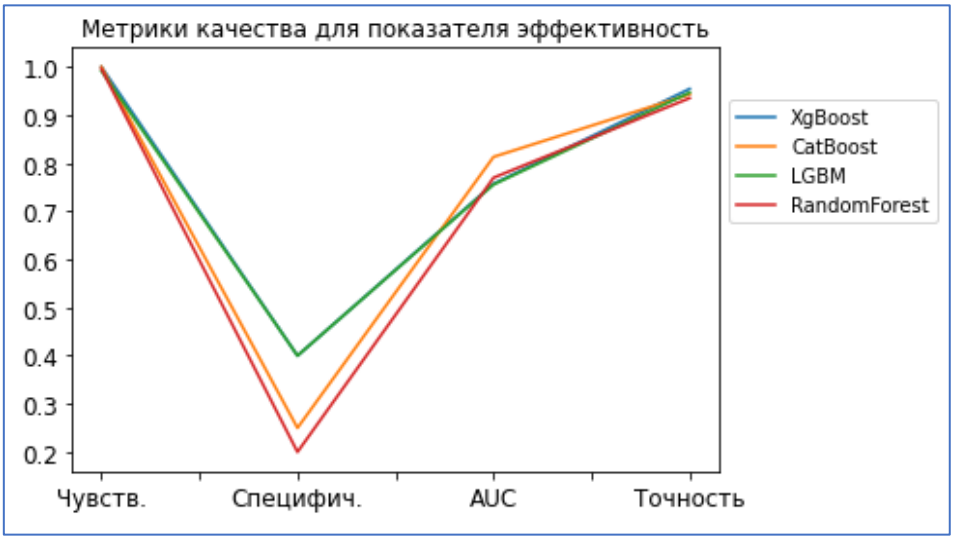

Рисунок 2 – Графики метрик качества
Figure 2 – Quality metrics graphs

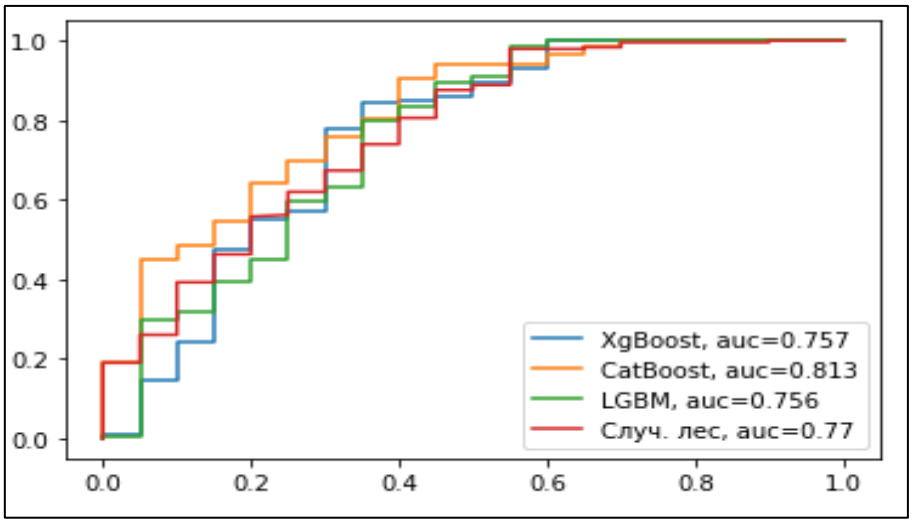

Рисунок 3 – Графики ROC-кривых
Figure 3 – ROC-curves

Таблица 4 – Значимость предикторов для модели «эффективность при выписке»
Table 4 – Significant predictors for the model of «efficiency in the statement»

XGBoost	*p-value*	*Случ. лес*	*p-value*	*CatBoost*	*p-value*	*LGBM*	*p-value*
наличие аритмии	0.0711	Тромбоцито-пения	0.0996	САД при поступл.	0.1867	аорта	0.1007
Аорта	0.0580	ПЖ	0.0818	PQ	0.1035	Лейкоцито-пения	0.0927
ИМТ	0.05	аорта	0.0793	СОЭ	0.0954	САД пост.	0.0921
пр. центр.д.	0.05	PQ	0.0675	Аорта	0.0774	PQ	0.0714

Затем более подробно были проанализированы признаки, влияющие на эффективность лечения. Визуализация некоторых результатов представлена на Рисунке 4. Исходя из Рисунка 4, можно сделать вывод, что аритмия присутствует примерно у 35% пациентов, у которых лечение оказалось неэффективным, и только у 15% пациентов, у которых лечение прошло успешно. Практически у всех 100% пациентов, у которых лечение оказалось неэффективным, в анамнезе были гипертонические кризы, тогда как среди пациентов, у которых оно было эффективным, таких только 60%. Неполная блокада ПНПГ (правой ножки пучка Гиса) в пять раз чаще встречается у пациентов, лечение которых оказалось неэффективным. Среди тех пациентов, которые до поступления в стационар принимали препараты центрального механизма действия (метилдопа, клонидин моксонидин, рилменидин) неудовлетворительный результат лечения в стационаре встречался в четыре раза чаще.

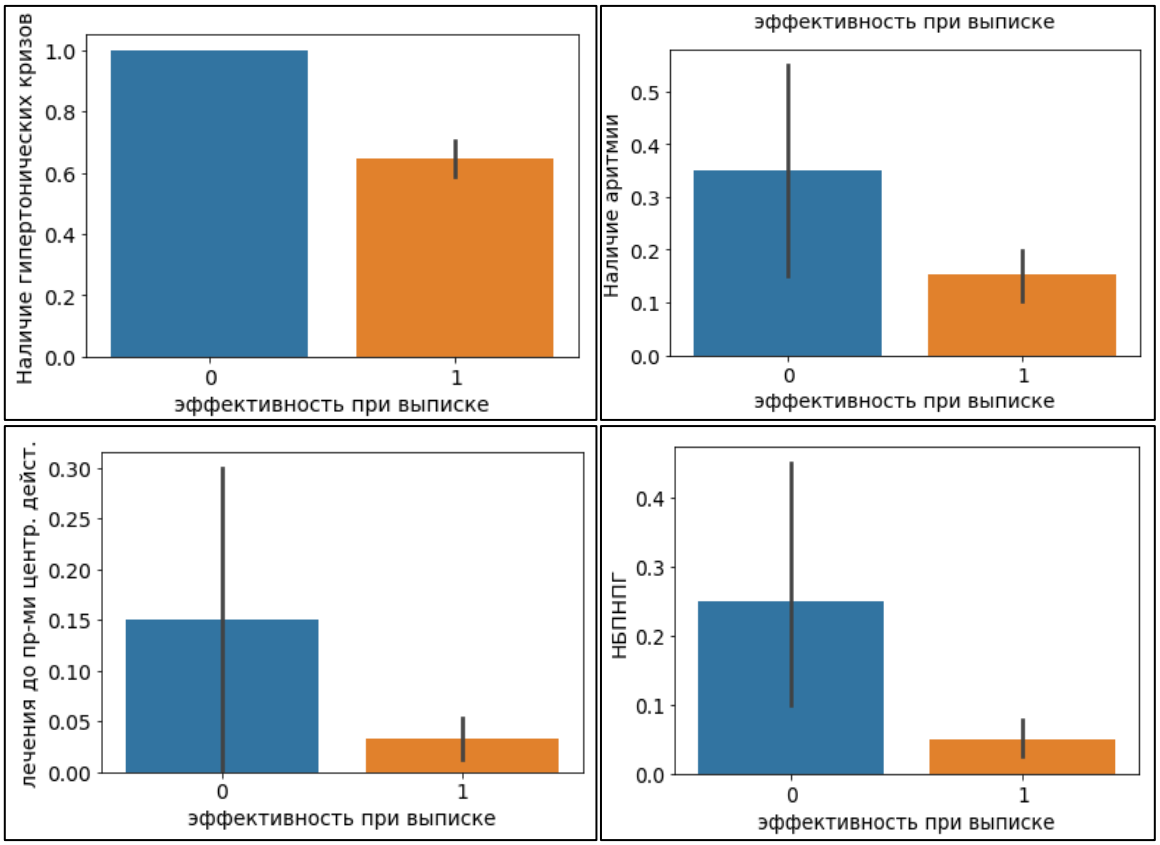

Рисунок 4 – Распределение клинических показателей в зависимости от эффективности лечения
Figure 4 – Distribution of clinical indicators depending on the effectiveness of treatment

3.2 Задача прогнозирования САД и ДАД при выписке

На следующем этапе задача прогнозирования эффективности лечения рассматривалась как задача регрессии. Для этого в качестве прогнозируемых переменных были выбраны показатели систолического и диастолического артериального давления пациента при выписке. Эта задача решалась с помощью регрессионных моделей машинного обучения: случайный лес, градиентный бустинг (XGBOOST, LIGHTGBM, CATBOOST),. В качестве входных параметров использовались те же значения, что описаны в первой задаче, но на выходе прогнозировалась значение САД и ДАД. Все модели были построены отдельно для значений САД и ДАД. На Рисунке 5 приведены столбчатые диаграммы значений R^2 для САД и ДАД при выписке построенных моделей.

Из графиков видно, что наибольшую точность показал метод градиентного бустинга – CATBOOST. Далее для анализа использовалась только эта модель. Параметры модели которые были заданы: глубина построенных деревьев (depth) равна 5, количество итераций (iterations) – 5000. Итоговая точность модели градиентного бустинга Catboost представлена в Таблице 5. Из результатов видно, что значение систолического артериального давления предсказывается лучше, чем диастолического. Средняя погрешность модели для САД и ДАД при выписке составляет 3.28 и 1.63 единиц соответственно. Если учитывать значения давления на 3-й день лечения, то качество прогноза можно немного улучшить.

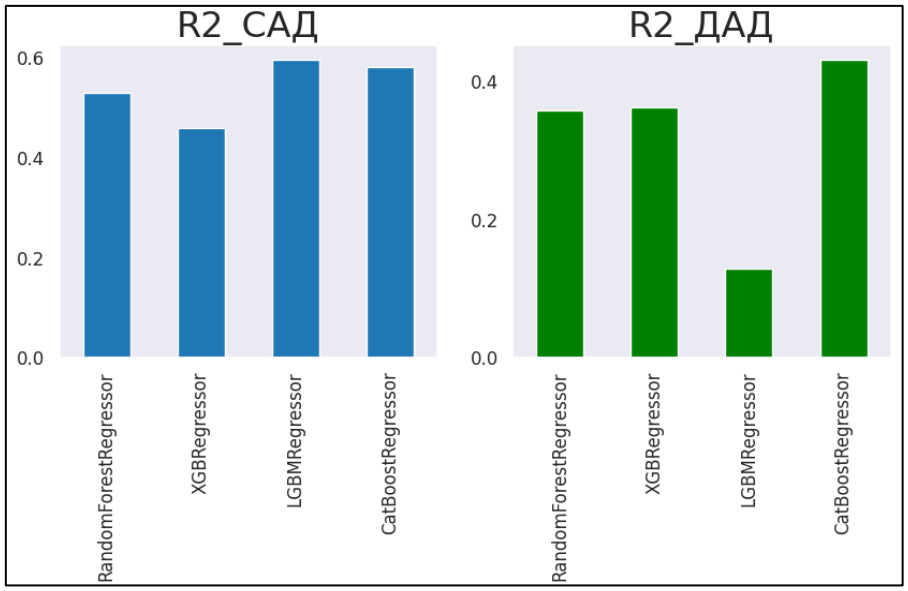

Рисунок 5 – Сравнение точности регрессионных моделей
Figure 5 – Comparison of the accuracy of regression models

Значения метрик качества с учетом показателей давления на третий день лечения в стационаре представлены также в Таблице 5.

Таблица 5 – Точность модели градиентный бустинг (Catboost)
Table 5 – Accuracy of Gradient Boosting (Catboost)

Прогнозируемое значение	R2	RMSE	MAE
САД при выписке	0.598	4.4	3.28
ДАД при выписке	0.4	2.5	1.63
САД при выписке (с учетом показателей на 3д)	0.6	4.3	3.24
ДАД при выписке (с учетом показателей на 3д)	0.4	2.5	1.61

На Рисунке 6 показаны наиболее значимые признаки для модели САД при выписке. В целом наибольшее влияние на выходной результат оказывают такие предикторы, как показатели САД и ДАД при поступлении, длительность АГ и QRS.

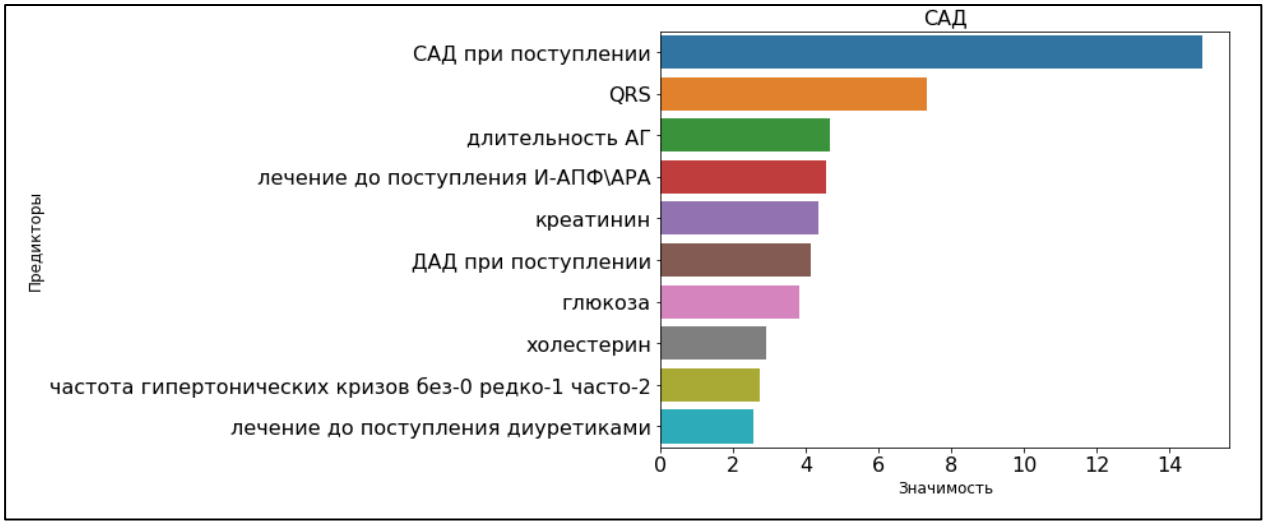

Рисунок 6 – Значимость предикторов (модель Catboost) при выписке
Figure 6 – Significance of predictors (Catboost model) at discharge

Одной из поставленных задач было оценить связь назначаемых препаратов с результатами артериального давления при выписке. На Рисунке 7 изображены кривые распределения значений САД при выписке для каждой исследуемой группы препаратов: 0 – препарат не назначался, 1 – назначался.

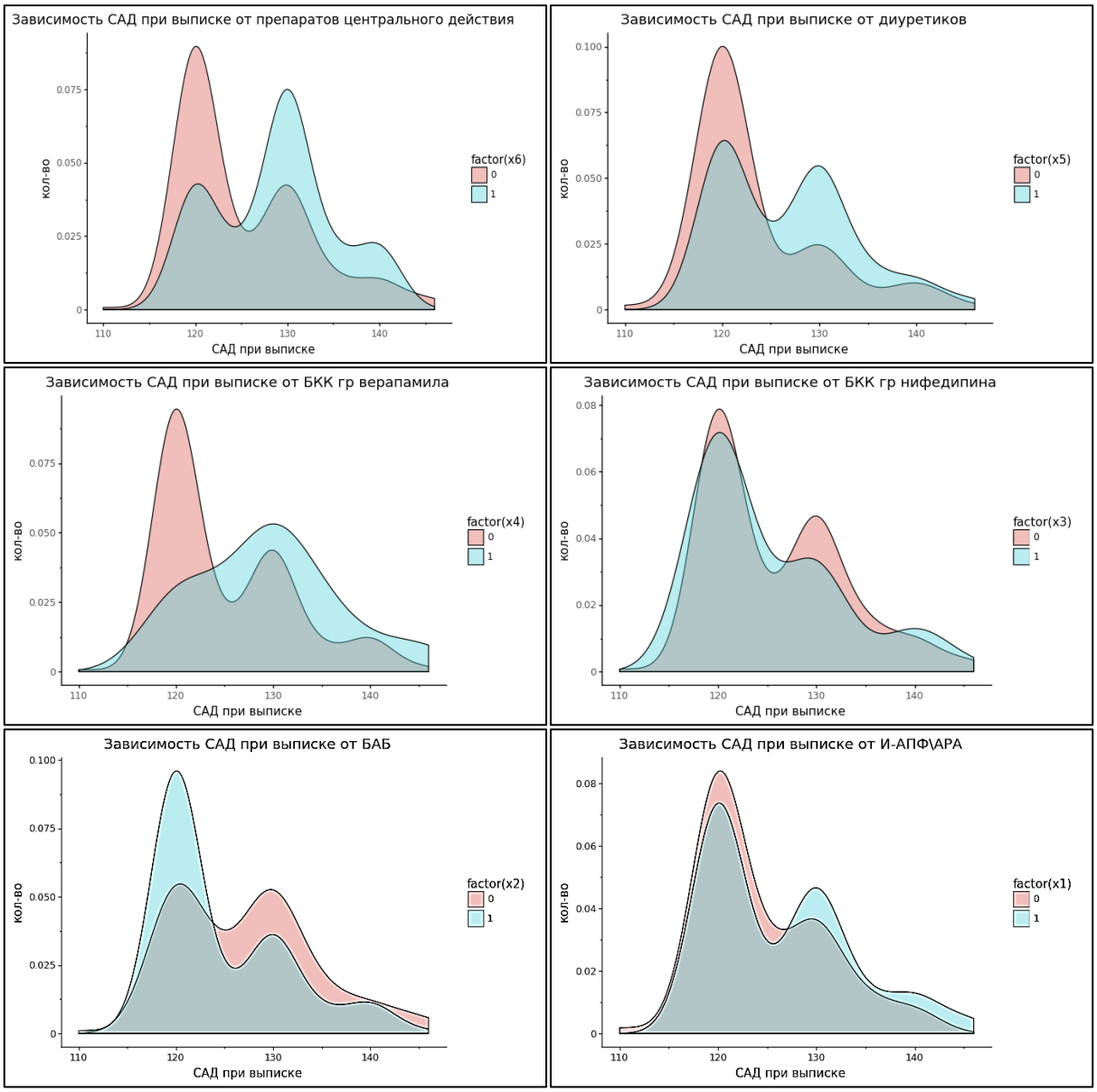

Рисунок 7 – Распределение САД при выписке в зависимости от применяемого препарата
Figure 7 – The distribution of SBP at discharge depending on the medicine used

Глядя на эти графики, можно отметить, что распределение значений САД при выписке существенно различается при назначении препаратов центрального действия, БАБ (бета-адреноблокаторов), БКК гр. верапамила и диуретиков и практически не различается при назначении И-АПФ\АРА и БКК гр. нифедипина. При этом только для БАБ среднее значение САД при выписке ниже у тех, пациентов, которые принимали этот препарат, чем у тех, которые не принимали. Для препаратов центрального действия, БКК гр. верапамила и диуретиков оно выше. Вероятно, это связано с тем, что назначение БАБ повышает среднюю эффективность терапии. Тем не менее, нельзя считать назначение

остальных препаратов неэффективным. На Рисунке 8 для сравнения приведены распределения САД при поступлении у пациентов, которым были назначены БАБ и И-АПФ\АРА. Видно, что последующая терапия оказала значимый эффект. Вместе с тем, нельзя не отметить, что наиболее «тяжелые хвосты» имеют распределения тех пациентов, которым были назначены БКК гр. верапамила. Именно к этой группе принадлежит 40% пациентов, лечение которых оказалось неэффективным.

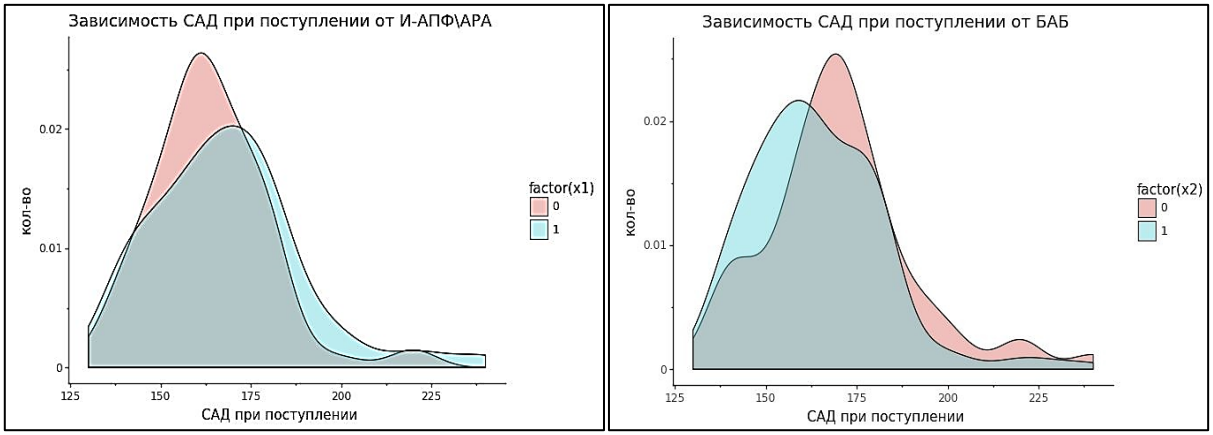

Рисунок 8 – Распределение САД при поступлении для двух групп назначаемых препаратов
Figure 8 – Distribution of SBP at admission for two groups of prescribed medicines

3.3 Задача построения моделей назначения отдельных препаратов

Следующим этапом данного исследования было необходимо на основе имеющихся клинических показателей пациента подобрать эффективное лечение. Задача решалась построением шести различных моделей машинного обучения (по каждой группе препаратов) для бинарной классификации – назначать данную группу препаратов или нет. Для этого в исходной выборке рассматривались только пациенты с успешным лечением, размер выборки составил 242 пациента.

Для каждого препарата были построены модели классификации с использованием четырех указанных ранее методов. Точность моделей, как и ранее, проверялась по кросс-проверке. Результаты качества построенных моделей приведены в Таблице 6.

Таблица 6 – Точность моделей классификации для каждого препарата
Table 6 – Accuracy of classification models for each medicine

	Чувствит	Специфич	AUC_ROC	Accuracy
БАБ				
XGBoost	0.78	0.58	0.73	0.7
CatBoost	0.92	0.58	0.77	0.76
LGBM	0.77	0.57	0.62	0.68
Случайный лес	0.81	0.54	0.71	0.69
И-АПФ\АРА				
XGBoost	0.81	0.46	0.68	0.70
CatBoost	0.88	0.43	0.72	0.73
LGBM	0.8	0.44	0.66	0.69
Случайный лес	0.82	0.43	0.65	0.70

Таблица 6 – Продолжение
Table 6 - Continuation

	Чувствит	Специфич	AUC_ROC	Accuracy
БКК группы нифедипина				
XGBoost	0.4	0.9	0.797	0.78
CatBoost	0.23	0.94	0.78	0.76
LGBM	0.38	0.88	0.78	0.76
Случайный лес	0.13	0.94	0.68	0.76
БКК группы верапамила				
XGBoost	0.62	0.97	0.93	0.93
CatBoost	0.55	0.96	0.91	0.91
LGBM	0.61	0.96	0.9	0.9
Случайный лес	0.2	0.97	0.85	0.87
Диуретики				
XGBoost	0.76	0.55	0.75	0.69
CatBoost	0.82	0.46	0.75	0.68
LGBM	0.76	0.55	0.77	0.68
Случайный лес	0.82	0.44	0.72	0.69
Препараты центрального действия				
XGBoost	0.5	0.97	0.878	0.91
CatBoost	0.43	0.99	0.83	0.93
LGBM	0.53	0.97	0.85	0.91
Случайный лес	0.4	0.99	0.75	0.91

По результатам, приведенным в таблице, можно сделать вывод, что в среднем наилучшие показатели точности оказались у модели XGBoost, поэтому данная модель была выбрана для дальнейшего анализа в этой задаче. На Рисунке 9 изображены графики ROC-кривых для модели XGBoost для каждого препарата. Из рисунка видно, что точнее всего удалось предсказать необходимость назначения БКК группы верапамила.

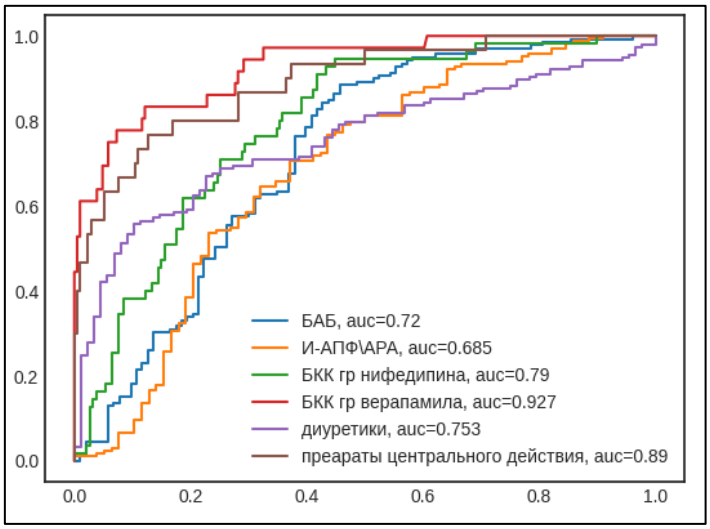

Рисунок 9 – Графики ROC-кривых для модели градиентный бустинг XGBoost
Figure 9 – ROC-curves for the XGBoost gradient boosting model

Значимыми признаками для модели XGBoost в задаче назначения отдельных препаратов оказались: наличие в анамнезе ХОБЛ, холестерин, ХСН, наличие осложнений сахарного диабета и показатель QRS.

3.4 Подход к выбору антигипертензивной терапии с использованием методов машинного обучения

Таким образом, в рамках исследования было построено 3 группы моделей:

1. Модели прогнозирования эффективности лечения при выписке для заданной комбинации назначаемых препаратов.

2а. Модели прогнозирования САД и ДАД при выписке для заданной комбинации назначаемых препаратов.

2б. Модели прогнозирования САД и ДАД при выписке для заданной комбинации назначаемых препаратов с учетом показателей САД и ДАД на третий день.

3. Модели прогнозирования назначения отдельных препаратов.

Весь подход к процессу выбору антигипертензивной терапии можно представить в виде схемы, отображенной на Рисунке 10.

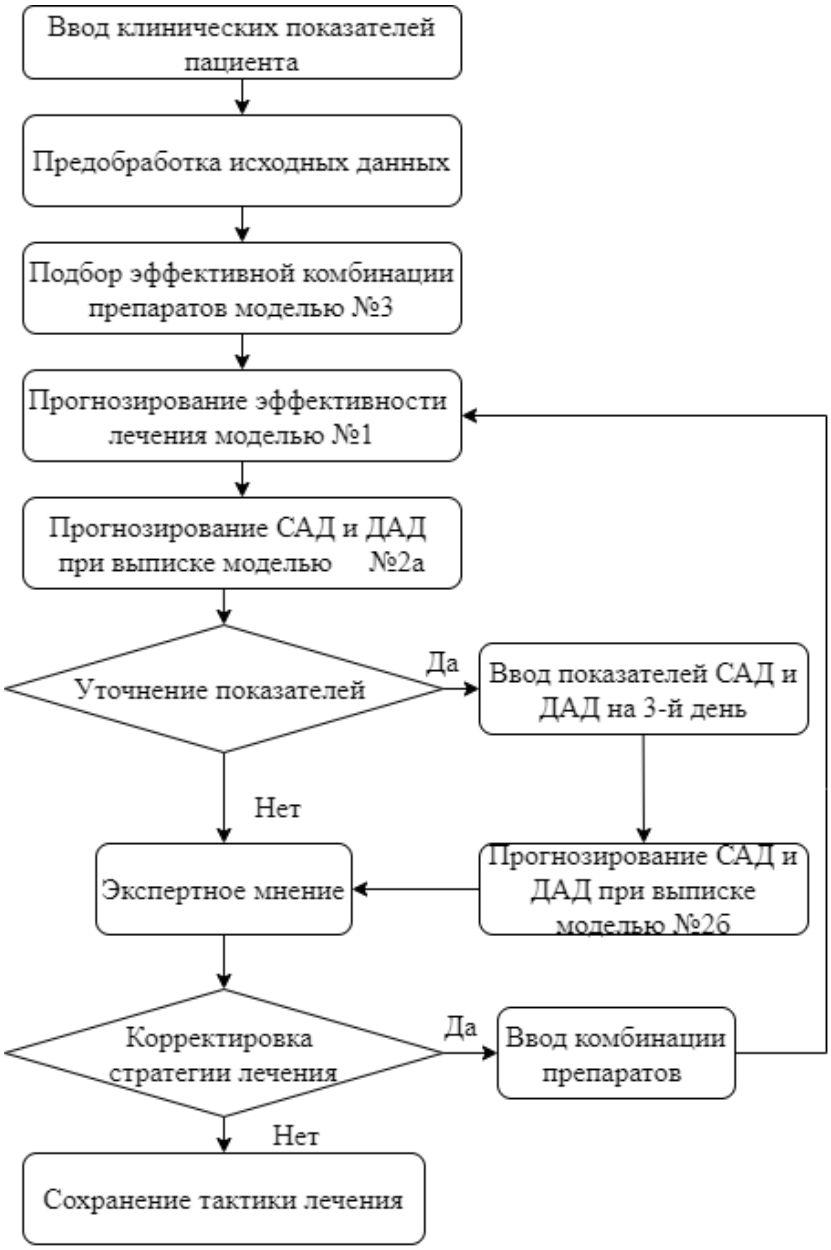

Рисунок 10 – Схема определения антигипертензивной терапии
Figure 10 – Antihypertensive therapy definition scheme

4 Заключение

В рамках исследования, описанного в данной статье на основе реальных данных пациентов кардиологического стационара с диагнозом артериальной гипертонии построен ряд моделей машинного обучения. На основе клинических показателей пациента и входного набора предполагаемой комбинации препаратов для приема в стационаре было спрогнозировано насколько лечение будет эффективно при выписке (точность *Accuracy* лучшей модели составила 0.95), и какими будут показатели систолического и диастолического артериального давления при выписке (точность R^2 лучшей модели составила 0.598). Также построены модели прогнозирования назначения отдельной группы препаратов (по шести группам препаратов) (точность *Accuracy* лучшей модели составила 0.93). Наибольшую точность показали модели градиентного бустинга XGBoost и Catboost для задачи классификации и регрессии соответственно.

По результатам исследования можно сделать выводы, что наиболее значимы для назначения лечения такие предикторы, как САД и ДАД при поступлении, СОЭ, PQ, размер аорты, креатинин и QRS.

Благодарности

Исследование выполнено при поддержке РФФИ, проект 20-37-90029 Аспиранты.

ЛИТЕРАТУРА

1. N. Ikeda. Control of hypertension with medication: a comparative analysis of national surveys in 20 countries. *Bulletin of the World Health Organization*. 2014;92(1):10-19.
2. S. Kjeldsen Updated national and international hypertension guidelines: a review of current recommendations. *Drugs*. 2014;6:2033-2051.
3. FeatureSelector: отбор признаков для машинного обучения на Python. Доступно по: https://proglib.io/p/feature-selector/ (дата обращения: 10.11.2020).
4. Will Koehrsen. A Feature Selection Tool for Machine Learning in Python : Towards Data Science. Доступно по: https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0/ (дата обращения: 01.11.2020).
5. Leo Breiman. Random Forests. *Machine Learning*. 2001:5-32.
6. The Ultimate Guide to Random Forest Regression. Доступно по: https://www.keboola.com/blog/random-forest-regression/ (дата обращения: 10.11.2020).
7. Шитиков В.К., Мастицкий С.Э. *Классификация, регрессия, алгоритмы Data Mining с использованием R*. Доступно по: https://github.com/ranalytics/data-mining/ (дата обращения: 09.09.2020)
8. Firyulina, M.A., Kashirina, I.L. Classification of cardiac arrhythmia using machine learning techniques. *Journal of Physics: Conference Series*, 2020;1479(1),012086. DOI: 10.1088/1742-6596/1479/1/012086
9. Каширина И.Л., Фирюлина М.А., Гафанович Е.Я. Анализ значимости предикторов выживаемости после инфаркта миокарда с помощью метода Каплана-Мейера. *Моделирование, оптимизация и информационные технологии.* 2019;1(24):7-20. DOI: 10.26102/2310-6018/2019.24.1.007
10. Гафанович Е.Я., Каширина И.Л. Модель-ориентированный подход к выбору антигипертензивной терапии. *Врач-аспирант.* 2015;(3.1):183-191.

REFERENCES

1. N. Ikeda. Control of hypertension with medication: a comparative analysis of national surveys in 20 countries. *Bulletin of the World Health Organization*. 2014;92(1):10-19.
2. S. Kjeldsen Updated national and international hypertension guidelines: a review of current recommendations. *Drugs*. 2014;6:2033-2051.
3. FeatureSelector: feature selection for machine learning in Python. Available at: https://proglib.io/p/feature-selector/ (accessed 10.11.2020).
4. Will Koehrsen. A Feature Selection Tool for Machine Learning in Python : Towards Data Science. Available at: https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0/ (accessed 01.11.2020).
5. Leo Breiman. Random Forests. *Machine Learning*. 2001:5-32.
6. The Ultimate Guide to Random Forest Regression. Available at: https://www.keboola.com/blog/random-forest-regression/ (accessed 10.11.2020).
7. Shitikov V.K., Mastitsky S.E. *Classification, regression, Data Mining algorithms using R*. Available at: https://github.com/ranalytics/data-mining/ (accessed 09.09.2020)
8. Firyulina, M.A., Kashirina, I.L. Classification of cardiac arrhythmia using machine learning techniques. *Journal of Physics: Conference Series*, 2020;1479(1),012086. DOI: 10.1088/1742-6596/1479/1/012086
9. Kashirina I. L., Firyulina M. A., Gafanovich E. Y. Analysis of the significance of predictors of survival after myocardial infarction using the Kaplan-Meyer method. *Modeling, optimization and information technology*. 2019;1(24):7-20. DOI: 10.26102/2310-6018/2019.24.1.007
10. Gafanovich E.YA., Kashirina I.L. Model-oriented approach to the choice of antihypertensive therapy. *Vrach-aspirant.* 2015;(3.1):183-191.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Фирюлина Мария Андреевна, аспирант, кафедра математических методов исследования операций, ФГБОУ ВО «Воронежский государственный университет», Воронеж, Российская Федерация.
e-mail: mashafiryulina@mail.ru
ORCID: 0000-0003-3468-5514

Mariya A. Firyulina, Phd Student, Mathematical Methods of Operations Research Department, Voronezh state university, Voronezh, Russian Federation.

Каширина Ирина Леонидовна, д.т.н., профессор, кафедра математических методов исследования операций, ФГБОУ ВО «Воронежский государственный университет», Воронеж, Российская Федерация.
e-mail: kash.irina@mail.ru
ORCID: 0000-0002-8664-9817

Irina L. Kashirina, Doctor of Technical Sciences, Professor, Mathematical Methods of Operations Research Department, Voronezh state university, Voronezh, Russian Federation.

Гафанович Елена Яковлевна, к. м. н., доцент кафедры факультетской терапии лечебного факультета, ФГБОУ ВО «Саратовский ГМУ им. В.И. Разумовского Минздрава России»
ORCID: 0000-0001-9122-6483

Elena Ya. Gafanovich, PhD, Associate Professor of the Department of Faculty Therapy of the Faculty of Medicine, Saratov State Medical University named after V. I. Razumovsky of the Ministry of Health of the Russian Federation.