УДК 519.2

© 2025 г. **М.З. Ермолицкая[1,2],** канд. биол. наук
[1](Институт автоматики и процессов управления ДВО РАН, Владивосток)
[2](Владивостокский государственный университет)

ПРОГНОЗИРОВАНИЕ НАЛИЧИЯ ХРОНИЧЕСКОЙ БОЛЕЗНИ ПОЧЕК У ПАЦИЕНТОВ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ*

В ходе проведения данного исследования были выделены показатели, медианные значения которых статистически значимо различаются в группах условно здоровых пациентов и пациентов с наличием хронической болезни почек. На основе выделенных показателей с помощью алгоритма дерева решений разработана классификационная модель, точность которой на тестовой выборке составила 0,941.
Ключевые слова: статистический анализ данных, логистическая регрессия, деревья решений, ROC-анализ, матрица неточностей.

DOI: 10.22250/18142400_2025_85_3_54

Введение

Хроническая болезнь почек (ХБП) представляет собой значимую проблему для здравоохранения. Ее распространенность среди населения России составляет 10–15% [1]. Заболеваемость сопровождается резким снижением качества жизни больных, потерей трудоспособности и повышенной смертностью. Ранняя диагностика, своевременное лечение ХБП позволяют замедлить прогрессирование заболевания и увеличить продолжительность жизни больных [2]. Для этого широко используются методы математической статистики и машинного обучения, открывающие новые горизонты для анализа сложных биологических данных и выявления закономерностей, незаметных при традиционных подходах. Статистические методы позволяют оценивать влияние различных факторов на здоровье человека, выявлять группы риска и прогнозировать развитие заболеваний [3 – 6]. Алгоритмы машинного обучения, такие как нейронные сети, деревья решений и метод опорных векторов, успешно применяются для диагностики заболеваний, выявления биомаркеров и предсказания эффективности лечения [7 – 9]. Сочетание методов математической статистически и машинного обучения позволяет создавать комплексные модели, способные решать широкий спектр задач в биомеди-

* Работа выполнена в рамках государственного задания FWFW-2022-0002.

цине. Интеграция этих методов становится ключевым фактором прогресса в современной медицине, открывая путь к более точной диагностике, эффективному лечению и профилактике заболеваний.

Целью данного исследования является разработка классификационной модели наличия хронической болезни почек у пациентов с помощью методов машинного обучения.

Материалы и методы

Исследованию подлежали лабораторные данные, полученные из сыворотки венозной крови, по 177 пациентам (36 человек с диагнозом хроническая болезнь почек и 141 человек с отсутствием такого диагноза) обоего пола в возрасте от 46 до 80 лет. Данные собраны на базе ФГБОУ ВО ТГМУ Минздрава России и ФГАОУ ВО ДВФУ Медицинский центр за период с 2017 по 2022 гг. Данные не полные. Участие в исследовании было основано на информированном согласии пациентов.

Статистический анализ данных осуществляли в программе RStudio (Version 1.0.136) с помощью непараметрических методов. Для определения различий медианных значений показателей в группах использовали критерий Крускела-Уоллиса (kruskal.test {stats}). Статистика критерия вычисляется по следующей формуле:

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1), \quad (1)$$

где n – объем выборки; R_i – ранг i-той выборки.

Различия считали достоверными на уровне значимости менее 0,05.

Для построения прогнозной модели использовали следующие методы: логистическую регрессию (glm {stats}) и деревья решений. Предварительно исходные данные были случайным образом разделены на обучающую и тестовую выборки (75% и 25% соответственно) с учетом групп (createDataPartition {caret}).

Логистическая функция имеет следующий вид:

$$P(y) = \frac{1}{1+e^{-y}}, \quad (2)$$
$$y = b_0 + b_1 x_1 + ... + b_n x_n,$$

где b_i – коэффициенты регрессии; n – количество предикторов; P – вероятность того, что объект относится к одной из двух групп.

В качестве методов деревьев решений использовали два алгоритма. Деревья условного вывода (ctree{partykit}) – непараметрический класс деревьев принятия решений**,** который основан на рекурсивном разбиении зависимой переменной с учетом значений корреляций. И алгоритм, заложенный в функции rpart {rpart}, позволяющий выполнять рекурсивный выбор при минимальной сумме квадратов внутригрупповых отклонений для всех узлов дерева с перекрестной проверкой. Каждый алгоритм делит исходный набор

данных на подмножества. Деление осуществляется на основе логических правил вида "Если …, то …".

Расчет производительности модели осуществляли с помощью ROC-анализа с построением матрицы неточностей (confusionMatrix {caret}). Точность модели (Accuracy) рассчитывалась как доля правильно классифицированных объектов (количество правильно классифицированных объектов к общему числу объектов); чувствительность (Sensitivity) - как процент верно предсказанных позитивных исходов; специфичность (Specificity) - процент верно предсказанных негативных исходов. Для визуальной оценки качества бинарной модели строили ROC-кривую (roc {pROC}), где площадь под кривой (AUC) является суммарной мерой точности прогноза [10]. который основан на рекурсивном разбиении зависимой переменной с учетом значений корреляций. И алгоритм, заложенный в функции rpart {rpart}, позволяющий выполнять рекурсивный выбор при минимальной сумме квадратов внутригрупповых отклонений для всех узлов дерева с перекрестной проверкой. Каждый алгоритм делит исходный набор данных на подмножества. Деление осуществляется на основе логических правил вида "Если …, то …".

Расчет производительности модели осуществляли с помощью ROC-анализа с построением матрицы неточностей (confusionMatrix {caret}). Точность модели (Accuracy) рассчитывалась как доля правильно классифицированных объектов (количество правильно классифицированных объектов к общему числу объектов); чувствительность (Sensitivity) - как процент верно предсказанных позитивных исходов; специфичность (Specificity) - процент верно предсказанных негативных исходов. Для визуальной оценки качества бинарной модели строили ROC-кривую (roc {pROC}), где площадь под кривой (AUC) является суммарной мерой точности прогноза [10].

Результаты

В ходе статистического анализа исходных данных были выявлены различия в группах ХБП по следующим семи показателям (табл. 1).

Таблица 1

Показатели	Медианные значения [Q1, Q3]		p-value
	ХБП=0	ХБП=1	
Возраст	65 [60; 68]	68 [64; 73,25]	0,0017
Креатинин	86 [74; 94,05]	116 [96;129,2]	3,02e-11
IL 13	37,08 [25,44; 65,28]	75,14 [74,99; 75,29]	0,0369
IL 6	0,16 [0,07; 0,51]	0,52 [0,16; 1,375]	0,0386
IL 4	2,93 [2,768; 5,358]	2,13 [1,51; 2,688]	0,0272
MMP 9	216,5 [129,2; 338,4]	306,8 [165,2; 395,1]	0,0291
TIMP 3	2,15 [1,19; 2,745]	3,86 [3,44; 4,56]	0,0209

Примечание. ХБП=0 – группа условно здоровых пациентов, ХБП=1 – группа пациентов с наличием ХБП.

Так как в данных имеются пропущенные значения, для сохранения большего числа наблюдений было решено в качестве предикторов для по-

строения модели использовать следующие показатели: возраст, креатинин, IL-6 и MMP-9. Медианные значения показателей статистически значимо различаются в группах пациентов с наличием ХБП и условно здоровых пациентов. Объем выборки составил 68 наблюдений (обучающая выборка – 51 наблюдение, тестовая выборка – 17).

На первом этапе для построения модели классификации воспользовались регрессионным анализом с логистическим видом зависимости. Согласно критерию Вальда полученная регрессия статистически значима (p-value = = 1,175e-06). Критерий Стьюдента показал, что только коэффициент регрессии при показателе креатинин значим (p-value = 2.1e-06), остальные коэффициенты регрессии незначимы (p-value >> 0,05). Такую модель нельзя использовать для прогнозирования.

На следующем этапе был применен метод деревьев решений. Результаты с применением алгоритма деревьев условного вывода (модель 1) показали, что на обучающей выборке 7 пациентов не верно классифицированы, на тестовой – 3 пациента. Показатели качества модели представлены в табл. 2.

Таблица 2

	Выборка	Точность / Accuracy	Чувствительность / Sensitivity	Специфичность/ Specificity
Модель 1	Обучающая выборка	0,863	0,837	1
	Тестовая выборка	0,824	0.786	1
Модель 2	Обучающая выборка	1	1	1
	Тестовая выборка	0,941	1	0,857

Согласно полученным оценкам качество модели 1 не высокое. На тестовой выборке процент верно предсказанных позитивных исходов равен лишь 78,6%, площадь под ROC-кривой составила 76,7% (рис. 1а).

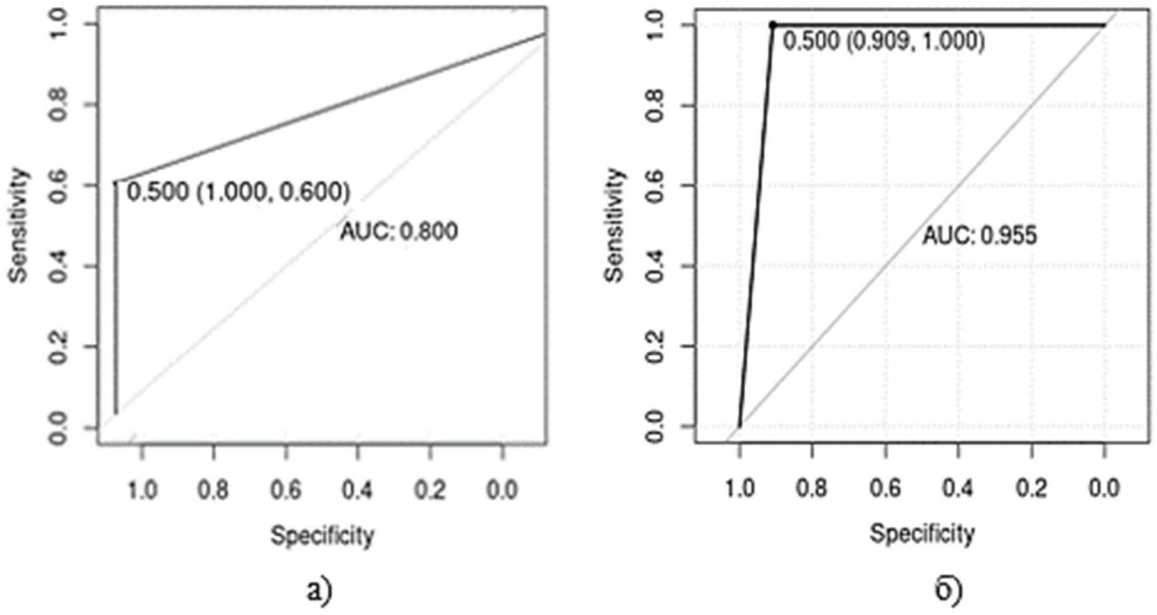

Рис. 1. График ROC-кривой на тестовой выборке по модели 1 (а) и модели 2(б).

Модель 2, построенная с помощь функции rpart{rpart}, показала лучшие результаты. На обучающей выборке все пациенты были верно классифицированы, на тестовой выборке имеется один ложно-положительный результат. Значение AUC на графике ROC-кривой на тестовой выборке составило 95,5 % (рис. 1б).

Структура полученного дерева решений по модели 2 представлена на рисунке 2, где каждый узел дерева содержит решающее правило разбиения. Показатели важности предикторов (varImp {caret}) в модели имеют следующие значения: креатинин – 17,135; возраст – 12,612; IL-6 – 11,673; MMP-9 – 9,347.

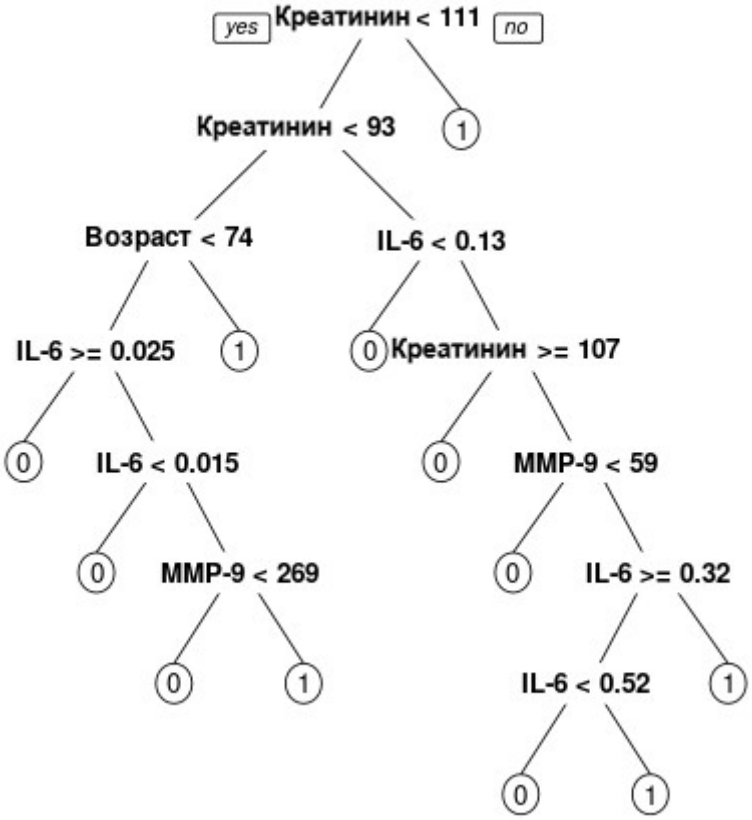

Рис. 2. Дерево решений модели 2.

Наименьшее влияние при классификации пациентов в модели 2 оказывает показатель MMP-9. Построили дерево решений без учета этого показателя. Тогда объем выборки увеличился и составил 79 наблюдений (60 наблюдений в обучающей выборке и 19 в тестовой выборке), но качество модели снизилось: на тестовой выборке чувствительность составила 0,857; специфичность – 0,8; AUC – 0,795.

Обсуждение

Для выявления у пациентов хронической болезни почек проводят измерения скорости клубочковой фильтрации (СКФ) и протеинурию для оценки общей функции почек. Эта процедура является трудоёмкой и дорогостоящей. В данном исследовании была предпринята попытка построения про-

гнозной модели на основе меньшего количества лабораторных данных с высокой чувствительностью и специфичностью.

С помощью алгоритма дерева решений в ходе перекрестной проверки, заложенных в функции rpart {rpart}, была разработана модель классификации пациентов на основе четырех показателей: креатинин, возраст, IL-6 и MMP-9. Точность модели (Accuracy) на тестовой выборке составила 0,941. Согласно полученным результатам, существенное влияние при прогнозировании ХБП оказывает креатинин и возраст пациентов, что подтверждается исследованиями ряда авторов [11 – 13]. В настоящее время известны различные подходы к расчету СКФ на основе креатинина для определения степени тяжести заболевания [14]. Разработанная нами модель позволяет лишь классифицировать пациента к одной из двух групп: условно здоровые пациенты и пациенты с наличием ХБП. Исследование проводилось на небольшом количестве наблюдений. Поэтому, для подтверждения качества данной модели необходимо провести ее анализ на новом наборе наблюдений.

Заключение

В ходе проведения данного исследования были выделены показатели, медианные значения которых статистически значимо различаются в группах условно здоровых пациентов и пациентов с наличием хронической болезни почек. На основе выделенных четырех показателей с помощью алгоритма дерева решений разработана классификационная модель, результаты тестирования которой демонстрируют достаточно высокое качество.

ЛИТЕРАТУРА

1. *Здравоохранение* в России. 2023: Стат.сб. / под ред. С.М. Окладников. – М.: Росстат, 2023.
2. *Есаян А.М., Арутюнов Г.П., Мелихов О.Г.* Распространенность хронической болезни почек среди пациентов, обратившихся в учреждения первичной медико-санитарной помощи. Результаты проспективного наблюдательного исследования в 12 регионах России // Клиническая нефрология. – 2021. – Т.13, № 3. – С. 6–16.
3. *Пырнова О.А.* Метод интеллектуального анализа данных для диагностики хронических заболеваний почек // Научно-технический вестник Поволжья. – 2023. – № 11. – С. 254–256.
4. *Заблоцкая О.В., Воробьева Е.П.* Ранние предикторы развития осложнений у пациентов с артериальной гипертензией: дисфункция почек, артериальная жесткость и нарушение углеводного обмена // Медицинские новости. – 2023. – № 4(21). – С. 38–40.
5. *Демчук О.В., Сукманова И.А.* прогнозирование частоты развития хронической болезни почек у пациентов с инфарктом миокарда и острым повреждением почек // Российский кардиологический журнал. – 2023. – Т. 28, № 6. – С. 24–30.
6. *Регрессионный* анализ клинических факторов риска прогрессирования ХБП у пожилых пациентов с СД 2 типа / Н.А. Первышин, С.В. Булгакова, А.А. Чертищева, Д.П. Курмаев и др. // Современные проблемы здравоохранения и медицинской статистики. – 2024. – № 1. – С. 241–255.

7. *Prediction* of renal transplantation outcome using artificial neural networks and investigating important risk factors / A. Zanghaei, Z. Rostami, A. Ameri, M. Salesi et al. // Urologiia. – 2023. – № 4. – pp. 82–89.
8. *Завьялова А.Н., Новикова В.П., Яковлева М.Н.* Саркопения у детей с детским церебральным параличем: факторы риска и критерии диагностики (пилотное исследование) // Профилактическая и клиническая медицина. – 2024. – № 1(90). – С. 14–23.
9. *Lovdal S.S., Den Hartigh R.J.R., Azzopardi G.* Injury prediction in competitive runners with machine learning // International journal of sports physiology and performance. – 2021. – Vol. 16(10). – pp. 1522–1531.
10. *Шитиков В.К., Мастицкий С.Э.* Классификация, регрессия и другие алгоритмы Data Mining с использованием R. – 2017. URL: https://github.com/ranalytics/data-mining (дата обращения: 04.03.2025).
11. *Creatinine* or cystatin C-based equations to estimate glomerular filtration in the general population: impact on the epidemiology of chronic kidney disease / P. Delanaye, E. Cavalier, O. Moranne, L. Lutteri et al. // BMC Nephrology. 2013. Mar 12;14:57. URL: https://bmcnephrol.biomedcentral.com/articles/10.1186/1471-2369-14-57 (дата обращения: 04.03.2025).
12. *Using* standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. / A.S. Levey, J. Coresh, T. Greene, L.A. Stevens et al. // Ann Intern Med. 2006. 145:247-254. URL: https://pubmed.ncbi.nlm.nih.gov/16908915/ (дата обращения: 04.03.2025).
13. *The definition*, classification, and prognosis of chronic kidney disease: a KDIGO controversies conference report / A.S. Levey, P.E. de Jong, J. Coresh, M. EI. Nahas et al. // Kidney Int. – 2011. – Vol. 80(1). – pp. 17–28.
14. *Креатинин* в современной оценке функционального состояния почек (обзор литературы и собственные данные) / И.Г. Каюков, О.В. Галкина, Е.И. Тимшина, И.М. Зубина и др. // Нефрология. – 2020. – Т. 24, № 4. – С. 21 – 36.

Статья представлена к публикации членом редколлегии А.И. Абакумовым.

E-mail:
Ермолицкая Марина Захаровна – ermmz@mail.ru.