

МЕТОДЫ ДИАГНОСТИКИ И ТЕХНОЛОГИИ

Научная статья

УДК 616-006.66

DOI: 10.17816/pmj423130-143

ВОЗМОЖНОСТЬ ПРЕДСКАЗАНИЯ ВЕРОЯТНОСТИ РЕЦИДИВА РАКА ЩИТОВИДНОЙ ЖЕЛЕЗЫ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

**М.А. Барулина¹, И.Ю. Бендик¹, И.И. Коваленко¹, М.А. Полиданов^{2*}, Р.П. Петрунькин²,
В.Н. Кудашкин³, К.А. Волков⁴, А.Р. Кравченко⁴, В.В. Масляков^{4,5}, С.В. Капралов⁴,
Г.Э. Асланов⁴, Е.В. Лосякова³, И.С. Обухов³, А.Д. Осина⁴, А.К. Курмаева⁴**

¹Пермский государственный национальный исследовательский университет,

²Университет «Реавиз», г. Санкт-Петербург,

³Самарский государственный медицинский университет,

⁴Саратовский государственный медицинский университет имени В.И. Разумовского,

⁵Медицинский университет «Реавиз», г. Саратов, Российская Федерация

© Барулина М.А., Бендик И.Ю., Коваленко И.И., Полиданов М.А., Петрунькин Р.П., Кудашкин В.Н., Волков К.А., Кравченко А.Р., Масляков В.В., Капралов С.В., Асланов Г.Э., Лосякова Е.В., Обухов И.С., Осина А.Д., Курмаева А.К., 2025
e-mail: maksim.polidanoff@yandex.ru

[Барулина М.А. – доктор физико-математических наук, директор Физико-математического института; ORCID: 0000-0003-3867-648X; Бендик И.Ю. – магистр I курса Физико-математического института; ORCID: 0009-0000-7851-9492; Коваленко И.И. – заведующий Центром искусственного интеллекта Физико-математического института; ORCID: 0000-0003-4450-1184; Полиданов М.А. (*контактное лицо) – советник Российской академии естествознания (РАЕ), специалист научно-исследовательского отдела, ассистент кафедры медико-биологических дисциплин, ORCID: 0000-0001-7538-7412; Петрунькин Р.П. – студент III курса лечебного факультета, ORCID: 0009-0003-3206-7920; Кудашкин В.Н. – врач-ординатор кафедры хирургии с курсом сердечно-сосудистой хирургии Института профессионального образования, ORCID: 0000-0001-9099-3517; Волков К.А. – студент III курса Института клинической медицины, ORCID: 0000-0002-3803-2644; Кравченко А.Р. – кандидат медицинских наук, доцент, доцент кафедры детских болезней лечебного факультета, ORCID: 0000-0003-2738-4510; Масляков В.В. – доктор медицинских наук, профессор, профессор кафедры мобилизационной подготовки здравоохранения и медицины катастроф, профессор кафедры хирургических болезней, ORCID: 0000-0001-6652-9140; Капралов С.В. – доктор медицинских наук, доцент, заведующий кафедрой факультетской хирургии и онкологии, ORCID: 0000-0001-5859-7928; Асланов Г.Э. – студент VI курса Института клинической медицины, ORCID: 0009-0009-9497-5725; Лосякова Е.В. – студент VI курса Института педиатрии, ORCID: 0009-0003-8286-4266; Обухов И.С. – студент VI курса Института педиатрии, ORCID: 0009-0007-5573-8431; Осина А.Д. – студентка VI курса Института клинической медицины, ORCID: 0009-0001-5294-3436; Курмаева А.К. – студентка VI курса Института клинической медицины, ORCID: 0009-0002-0886-6290].

POSSIBILITY OF PREDICTING THE PROBABILITY OF THYROID CANCER RECURRENCE BY MACHINE LEARNING METHODS

M.A. Barulina¹, I.Yu. Bendik¹, I.I. Kovalenko¹, M.A. Polidanov^{2}, R.P. Petrunkin², V.N. Kudashkin³, K.A. Volkov⁴, A.R. Kravchenya⁴, V.V. Maslyakov^{4,5}, S.V. Kapralov⁴, H.E. Aslanov⁴, Ye.V. Losyakova³, I.S. Obukhov³, A.D. Osina⁴, A.K. Kurmaeva⁴*

¹Perm State National Research University,

²University «Reaviz», Saint Petersburg,

³Samara State Medical University

⁴Saratov State Medical University named after V.I. Razumovsky

⁵Medical University «Reaviz», Saratov, Russian Federation

Цель. Разработка модели машинного обучения по предсказанию факта рецидива у пациентов с раком щитовидной железы после проведенного оперативного вмешательства.

Материалы и методы. В соответствии с целью исследования были проанализированы истории болезни 300 пациентов с выполненным оперативным вмешательством по поводу рака щитовидной железы. Средний возраст – 43,54 года. Всем включенным в исследование больным было проведено комплексное обследование согласно клиническим рекомендациям по диагностике и лечению больных РЩЖ. Выбор наиболее подходящей модели в машинном обучении критически важен, так как он напрямую влияет на точность и эффективность предсказания. Отбор лучшей модели был произведен через сравнение производительности различных алгоритмов на одной и той же обучающей выборке с использованием кросс-валидации. Каждая модель оценивалась по метрикам, таким как средняя точность и стандартное отклонение, что позволяет определить, какая из них демонстрирует наилучшие результаты. Лучше всего по показателю средней точности выявила себя модель случайного леса, она же в дальнейшем и использовалась. Обучение модели было произведено по матрице заранее определенных признаков. Используя параметрическую сетку (param_grid), можно эффективно настраивать гиперпараметры, такие как количество деревьев, максимальная глубина и минимальное количество

© Barulina M.A., Bendik I.Yu., Kovalenko I.I., Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Maslyakov V.V., Kapralov S.V., Aslanov H.E., Losyakova Ye.V., Obukhov I.S., Osina A.D., Kurmaeva A.K., 2025
e-mail: maksim.polidanov@yandex.ru

[Barulina M.A. – DSc (Physics and Mathematics), Director of the Institute of Physics and Mathematics, ORCID: 0000-0003-3867-648X; Bendik I.Yu. – 1st-year Master's Student of the Institute of Physics and Mathematics, ORCID: 0009-0000-7851-9492; Kovalenko I.I. – Head of the Center for Artificial Intelligence of the Institute of Physics and Mathematics, ORCID: 0000-0003-4450-1184; Polidanov M.A. (*contact person) – Advisor to the Russian Academy of Natural Sciences (RANS), Research Department Specialist, Assistant of the Department of Biomedical Disciplines, ORCID: 0000-0001-7538-7412; Petrunkin R.P. – 3rd-year Student of the Faculty of Medicine, ORCID: 0009-0003-3206-7920; Kudashkin V.N. – Resident of the Department of Surgery with a Course in Cardiovascular Surgery of the Institute of Professional Education, ORCID: 0000-0001-9099-3517; Volkov K.A. – 3rd-year Student of the Institute of Clinical Medicine, ORCID: 0000-0002-3803-2644; Kravchenya A.R. – PhD (Medicine), Associate Professor, Associate Professor of the Department of Pediatric Diseases of the Faculty of Medicine, ORCID: 0000-0003-2738-4510; Maslyakov V.V. – DSc (Medicine), Professor, Professor of the Department of Mobilization Preparation of Healthcare and Disaster Medicine, Professor of the Department of Surgical Diseases, ORCID: 0000-0001-6652-9140; Kapralov S.V. – DSc (Medicine), Associate Professor, Head of the Department of Faculty Surgery and Oncology, ORCID: 0000-0001-5859-7928; Aslanov H.E. – 6th-year Student of the Institute of Clinical Medicine, ORCID: 0009-0009-9497-5725; Losyakova Ye.V. – 6th-year Student of the Institute of Pediatrics, ORCID: 0009-0003-8286-4266; Obukhov I.S. – 6th-year Student of the Institute of Pediatrics, ORCID: 0009-0007-5573-8431; Osina A.D. – 6th-year Student of the Institute of Clinical Medicine, ORCID: 0009-0001-5294-3436; Kurmaeva A.K. – 6th-year Student of the Institute of Clinical Medicine, ORCID: 0009-0002-0886-6290].

образцов для разделения, что поможет найти оптимальные настройки для нашей задачи. Для подбора гиперпараметров использовался метод RandomizedSearchCV. В процессе поиска гиперпараметров модель обучалась на тренировочных данных, отобранных как 70 % от исходного датасета. Итогом поиска определились следующие лучшие гиперпараметры для модели случайного леса для конкретно наших данных: `n_estimators = 161`; `min_samples_split = 5`; `max_leaf_nodes = 39`; `max_depth = 12`; `bootstrap = True`.

Результаты. В ходе исследования была обучена модель, которая продемонстрировала высокую точность целевого признака. Доля пациентов с послеоперационным рецидивом, правильно идентифицированных моделью, составила 98 % от общего числа пациентов с рецидивом, а доля пациентов без рецидива, верно классифицированных моделью «как пациенты, не имеющие риска рецидива», – 95 % от всех пациентов без рецидива. Это свидетельствует, что разработанная модель эффективно справляется с задачей классификации на основе медицинских параметров, что может быть особенно важно для принятия решений в клинической практике. Высокая точность указывает на надежность модели и ее способность правильно идентифицировать случаи рецидива, что может способствовать улучшению диагностики и лечения.

Выводы. В рамках исследования была разработана модель машинного обучения для предсказания высокой вероятности рецидива рака щитовидной железы на основе анализа медицинских параметров. Процесс разработки начался с тщательной предобработки данных, что является критически важным этапом в построении надежных моделей. В ходе предобработки были удалены выбросы и столбцы, содержащие однообразные значения, что позволило улучшить качество данных и избежать искажений в обучении модели. Также была проведена кодировка категориальных переменных, что обеспечило возможность их корректного использования в алгоритмах машинного обучения, и исключены коррелирующие признаки, чтобы минимизировать мультиколлинеарность и повысить интерпретируемость модели.

Ключевые слова. Рак щитовидной железы, щитовидная железа, машинное обучение, прогнозирование рецидивов, случайный лес, python.

Objective. To develop a machine learning model for predicting the fact of recurrence in patients with thyroid cancer after surgical intervention.

Materials and Methods. According to the aim of the study, the case histories of 300 patients who had undergone surgical intervention for thyroid cancer were analyzed. The average age was 43.54 years. All patients included in the study underwent a comprehensive examination according to the clinical recommendations on the diagnosis and treatment of patients with thyroid cancer. Selection of the most appropriate model in machine learning is critical as it directly affects the accuracy and efficiency of prediction. Selection of the best model was done through comparing the performance of different algorithms on the same training sample using cross-validation. Each model was evaluated on such metrics as average accuracy and standard deviation to determine which model demonstrates the best results. The random forest model performed best in terms of average accuracy and was used hereafter. The model was trained using a matrix of predefined features. Using param grid, we can efficiently adjust hyperparameters such as the number of trees, maximum depth and minimum number of samples for separation, which will help us to find the optimal settings for our task. RandomizedSearchCV method was used to select the hyperparameters. During the hyperparameter search process, the model was trained on training data selected as 70% of the original dataset. The search resulted in the following best hyperparameters for the random forest model for our data specifically: `n_estimators = 161`; `min_samples_split = 5`; `max_leaf_nodes = 39`; `max_depth = 12`; `bootstrap = True`.

Results. A model that demonstrated high target feature accuracy was trained during the study. The proportion of patients with postoperative recurrence correctly identified by the model was 98 % of all patients with recurrence, and the proportion of patients without recurrence correctly classified by the model «as patients at no risk of recurrence» was 95 % of all patients without recurrence. This shows that the developed model effectively handles the task of classification based on medical parameters, which may be particularly important for decision making in clinical practice. The high accuracy indicates the reliability of the model and its ability to identify cases of recurrence correctly, this may contribute to the improvement of diagnostics and treatment.

Conclusions. A machine learning model to predict a high probability of thyroid cancer recurrence based on the analysis of medical parameters was developed while carrying out the study. The development process be-

gan with careful data preprocessing, which is a critical step in reliable models' construction. During preprocessing, outliers and columns containing monotonic values were removed to improve the data quality and avoid distortions in the model training. Categorical variables were also coded to ensure that they could be used correctly in machine learning algorithms, and correlated features were excluded to minimize multicollinearity and increase the interpretability of the model.

Keywords. Thyroid cancer, thyroid, machine learning, recurrence prediction, random forest, python.

ВВЕДЕНИЕ

Рак щитовидной железы (РЩЖ) является одним из наиболее распространенных видов рака среди эндокринных заболеваний [1–4]. Несмотря на высокие показатели выживаемости при раннем обнаружении и адекватном лечении, проблема рецидива остается актуальной и требует особого внимания [5–7]. Рецидив заболевания может произойти даже после успешного лечения, что делает необходимым регулярное наблюдение за пациентами [8–11]. Однако предсказание рецидива на основе клинических показателей представляет собой сложную задачу для медицины в целом.

Специалисты в области онкологии сталкиваются с несколькими трудностями при оценке риска рецидива заболевания. Во-первых, клинические данные о пациентах могут быть разнообразными и многогранными, включая возраст, пол, степень дифференцировки опухоли, наличие метастазов и результаты предыдущих анализов. Эти факторы могут взаимодействовать друг с другом сложным образом, что затрудняет их интерпретацию. Во-вторых, традиционные методы оценки риска часто основаны на субъективных оценках врачей, что может привести к вариативности в диагнозах и рекомендациях по лечению.

Кроме того, время, затрачиваемое на анализ данных и принятие решений, может быть значительным. В условиях ограниченных ресурсов и увеличивающейся нагрузки на медицинских специалистов важно оптимизировать процесс диагностики и монито-

ринга пациентов. Неправильная оценка риска рецидива может привести не только к ухудшению состояния пациента, но и к ненужным затратам на дополнительные обследования и лечение.

В связи с вышеописанными трудностями возникает необходимость в разработке автоматизированной модели предсказания рецидива заболевания РЩЖ. Использование методов машинного обучения позволяет анализировать большие объемы данных и выявлять скрытые закономерности между различными клиническими показателями и вероятностью рецидива. Модель может быть обучена на исторических данных о пациентах, что позволит ей делать более точные прогнозы на основе новых входных данных.

Разработка такой модели не только повысит точность предсказаний, но и значительно сократит время, необходимое для анализа данных, что позволит врачам сосредоточиться на более важных аспектах лечения пациентов и улучшить качество медицинской помощи. Кроме того, автоматизация процесса оценки риска рецидива может привести к экономии средств как для медицинских учреждений, так и для пациентов.

Таким образом, создание модели предсказания рецидива заболевания РЩЖ представляет собой важный шаг к улучшению диагностики и лечения пациентов.

Цель исследования – разработка модели машинного обучения по предсказанию факта рецидива у пациентов с раком щитовидной железы после проведенного оперативного вмешательства.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

В соответствии с целью исследования были проанализированы истории болезни 300 пациентов с выполненным оперативным вмешательством по поводу РЩЖ. Средний возраст – 43,54 года. Всем включенным в исследование было проведено комплексное обследование согласно клиническим рекомендациям [12] по диагностике и лечению больных РЩЖ. Исходя из комплекса результатов обследования, отобранные пациенты, соответствующие следующим критериям включения: больные РЩЖ без подтвержденных метастазов со стадией заболевания от T1N0M0 до T3N0M0; отсутствие предшествующего и сопутствующего специального лечения (иммунотерапия или таргетная терапия); наличие информированного согласия на проводимое оперативное вмешательство и участие в исследовании.

Выбор наиболее подходящей модели в машинном обучении критически важен, так как он напрямую влияет на точность и эффективность предсказаний¹. Правильная модель позволяет лучше выявлять закономерности в данных и адаптироваться к специфике задачи. Неподходящая модель может привести к низкой производительности, «переобучению» или «недообучению», что затруднит интерпретацию результатов и принятие решений.

Было рассмотрено несколько самых распространенных:

- Logistic Regression (LR);
- Linear Discriminant Analysis (LDA);
- K-Nearest Neighbors (KNN);
- Classification and Regression Trees (CART);
- Gaussian Naive Bayes (NB);
- Support Vector Machines (SVM);
- Random Forest Classifier (RF).

Отбор лучшей модели был произведен через сравнение производительности различных алгоритмов на одной и той же обучающей выборке с использованием кросс-валидации. Каждая модель оценивалась по метрикам, таким как средняя точность и стандартное отклонение, что позволяло определить, какая из них демонстрирует наилучшие результаты (табл. 1).

Лучше всего по показателю средней точности выявила себя модель случайного леса, она же в дальнейшем и использовалась.

Обучение модели было произведено по матрице заранее определенных признаков, так как это позволяет систематически исследовать влияние различных параметров на производительность модели. Используя параметрическую сетку (`param_grid`), можно эффективно настраивать гиперпараметры², такие как количество деревьев, максимальная глубина и минимальное количество образцов для разделения, что поможет найти оптимальные настройки для нашей задачи.

Для подбора гиперпараметров использовался метод `RandomizedSearchCV`. Его особенность в том, что вместо того, чтобы проверять все возможные комбинации этих гиперпараметров, `RandomizedSearchCV`

¹ Предварительная обработка данных в машинном обучении: инструкция, инструменты, полезные ресурсы для начинающих, available at: <https://habr.com/ru/companies/skillfactory/articles/848858/>; A Simple Guide to Data Preprocessing in Machine Learning, available at <https://www.v7labs.com/blog/data-preprocessing-guide/>; Обзор методов классификации в машинном обучении с помощью Scikit-Learn, available at: <https://tproger.ru/translations/scikit-learn-in-python>

² Подбор гиперпараметров, available at: <https://education.yandex.ru/handbook/ml/article/podbor-giperparametrov/>; Гиперпараметрический поиск и оптимизация моделей, available at: <https://habr.com/ru/companies/otus/articles/754402/>; Метрики качества моделей бинарной классификации, available at: <https://loginom.ru/blog/classification-quality/>; Оценка качества в задачах классификации и регрессии, available at: <https://neerc.ifmo.ru/wiki/index.php?title>

Таблица 1

Результаты работы моделей

Название модели	Accuracy	Loss
Logistic Regression (LR)	0,894737	0,074432
Linear Discriminant Analysis (LDA)	0,884211	0,077352
K-Nearest Neighbors (KNN)	0,563158	0,094297
Classification and Regression Trees (CART)	0,921053	0,053931
Gaussian Naive Bayes (NB)	0,836842	0,086322
Support Vector Machines (SVM)	0,552632	0,026316
Random Forest Classifier (RF)	0,942105	0,043719

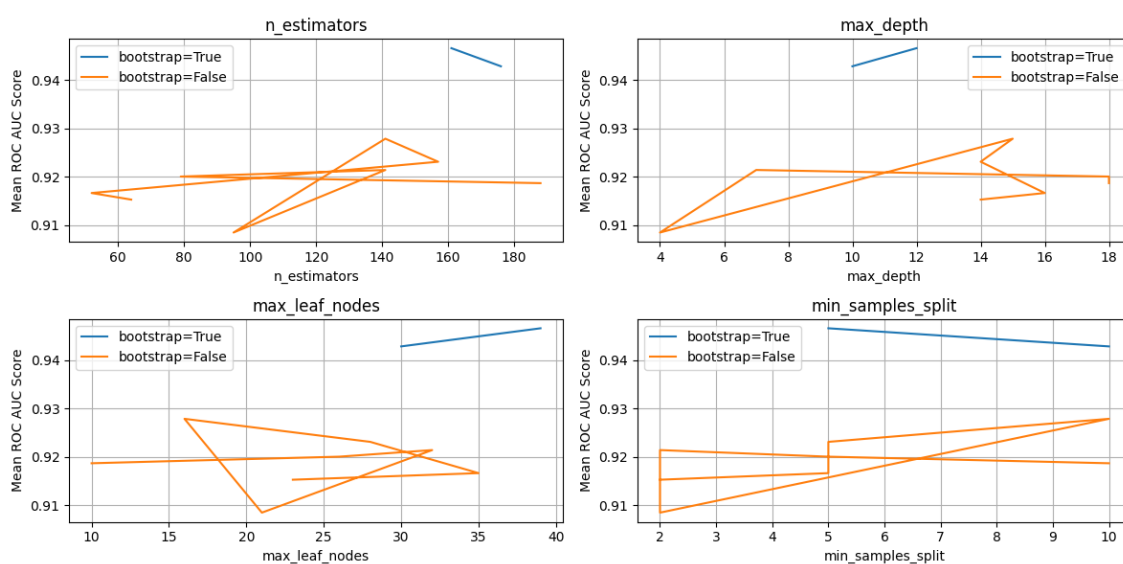


Рис. 1. Поиск гиперпараметров и обучение модели

случайным образом выбирает заданное количество комбинаций, что позволяет быстрее находить оптимальные настройки и особенно полезно при большом количестве гиперпараметров или их значений, так как помогает избежать чрезмерных затрат ресурсов и времени на обучение моделей. В процессе поиска этих гиперпараметров модель и обучается на тренировочных данных, отобранных как 70 % от исходного датасета, что впоследствии задается при вызове функции для обучения. Визуализация поиска и обучения изображена на рис. 1.

В поиске определились следующие лучшие гиперпараметры для модели случайного леса для конкретно наших данных:

- `n_estimators = 161`. Это количество деревьев в случайном лесу. Больше количество деревьев обычно приводит к более точным предсказаниям, так как модель становится более устойчивой к шуму в данных;
- `min_samples_split = 5`. Минимальное количество образцов, необходимых для разделения узла дерева. Этот параметр контролирует, насколько «глубоко» может расти дерево; более высокие значения предот-

вращают переобучение, снижая сложность модели;

- `max_leaf_nodes = 39`. Максимальное количество листовых узлов в дереве. Ограничение на количество листьев помогает контролировать сложность модели и может улучшить обобщающую способность;

- `max_depth = 12`. Максимальная глубина дерева. Этот параметр ограничивает, насколько глубоко может расти дерево, что также помогает предотвратить переобучение;

- `bootstrap = True`. Указывает, используется ли метод бутстрепа для создания подвыборок из данных при обучении каждого дерева. При установке этого параметра в `True` каждое дерево обучается на случайной подвыборке данных, что способствует раз-

нообразию деревьев и повышает общую производительность модели.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для анализа в среде разработки в предоставленном наборе данных были изменены названия столбцов на более короткие, латинскими буквами по следующей легенде (табл. 2).

Целевым признаком, для которого необходимо было разработать предиктивную модель, являлся признак «Послеоперационного рецидива» (в табл. 2 представлен как `pr`). Для этого признака было проверено распределение классов, оно представлено в виде круговой диаграммы (рис. 2).

Таблица 2

Легенда переименования столбцов

Новое название	Старое название
<code>id</code>	Код пациента
<code>age</code>	Возраст
<code>dotdm</code>	Длительность заболевания, месяц
<code>dbtt</code>	Диагноз по TNM. T
<code>dbtn</code>	Диагноз по TNM. N
<code>dbtm</code>	Диагноз по TNM. M
<code>ap</code>	Щелочная фосфатаза
<code>tc</code>	Кальций общий
<code>ttg</code>	ТТГ
<code>tfpl</code>	T _с свободный, пмоль/л
<code>cpm</code>	Кальцитонин, пг/мл
<code>phpl</code>	Паратгормон, пмоль/л
<code>at</code>	Антитела к тиреоглобулину, МЕ/мл
<code>rea</code>	РЭА, нг/мл
<code>cc</code>	Цитологическая классификация после ТАБ по системе Bethesda (диагностическая категория от 1 до 5)
<code>cd</code>	Сопутствующие заболевания
<code>cds</code>	Сопутствующие заболевания ССС
<code>cdg</code>	Сопутствующие заболевания ЖКТ
<code>cdd</code>	Сопутствующие заболевания ДС
<code>td</code>	Вид операции
<code>apas</code>	Щелочная фосфатаза после операции
<code>tcas</code>	Кальций общий после операции
<code>tao</code>	ТТГ после операции
<code>t4ao</code>	T _с после операции
<code>cas</code>	Кальцитонин после операции
<code>phas</code>	Паратгормон после операции

Окончание табл. 2

Новое название	Старое название
atas	Антитела к тиреоглобулину после операции
ras	РЭА после операции
dtah	Диагноз по TNM после гистологии. T
dtahN	Диагноз по TNM после гистологии. N
dtahM	Диагноз по TNM после гистологии. M
hsas	Время нахождения в стационаре после операции, дни
ic	Интраоперационные осложнения
pr	Послеоперационный рецидив

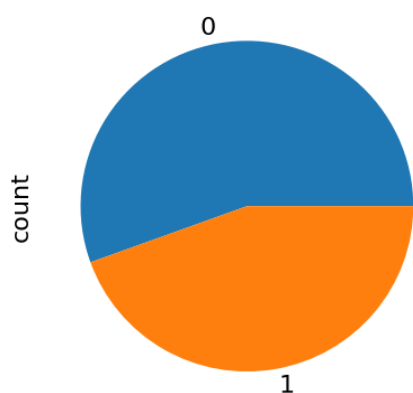


Рис. 2. Распределения классов в целевом признаке

Дисбаланс классов при этом не наблюдался. В связи с чем в дальнейшем с помощью метода corr, который вычисляет коэффициент корреляции Пирсона, являющийся мерой линейной зависимости между двумя переменными, из библиотеки Pandas для языка программирования Python³ в среде разработки Visual Studio Code была рассчитана корреляция всех признаков (без учета категориальных) между собой. Матрица корреляции признаков представлена на рис. 3.

По матрице видно, что есть два признака («cas» и «срт») сильно коррелирующих между собой. Для избегания избыточности информации и, как следствие, ухудшения обобщающей способности предсказательной модели было произведено удаление таких признаков. Был использован метод вычисления коэффициента корреляции для выявления высокоррелирующих признаков. Сначала создается матрица корреляции, после чего производится итерация по ее элементам для выявления пар признаков с абсолютным значением корреляции выше заданного порога (в данном случае 0,75). Все такие признаки добавляются в множество, которое затем используется для удаления этих признаков из исходного набора данных. В нашем случае признак всего один, а также заметна сильная корреляция признака «сс» с целевым признаком. Относительно него далее была составлена диаграмма рассеивания признаков с явным разделением на интересные классы.

С помощью функции sns.boxplot (df) из библиотеки Seaborn была представлена визуализация распределения данных в виде коробчатой диаграммы (boxplot). На ней можно сравнивать распределения различных групп данных, выявлять выбросы и оценивать вариативность. Видно отображение медианы (второй квартиль), первого (Q_1) и третьего (Q_3) квартиля, которые формируют межквартильный размах (IQR).

³ AI with Python – Supervised Learning: Classification, available at: https://www.tutorialspoint.com/artificial_intelligence_with_python/artificial_intelligence_with_python_supervised_learning_classification.htm

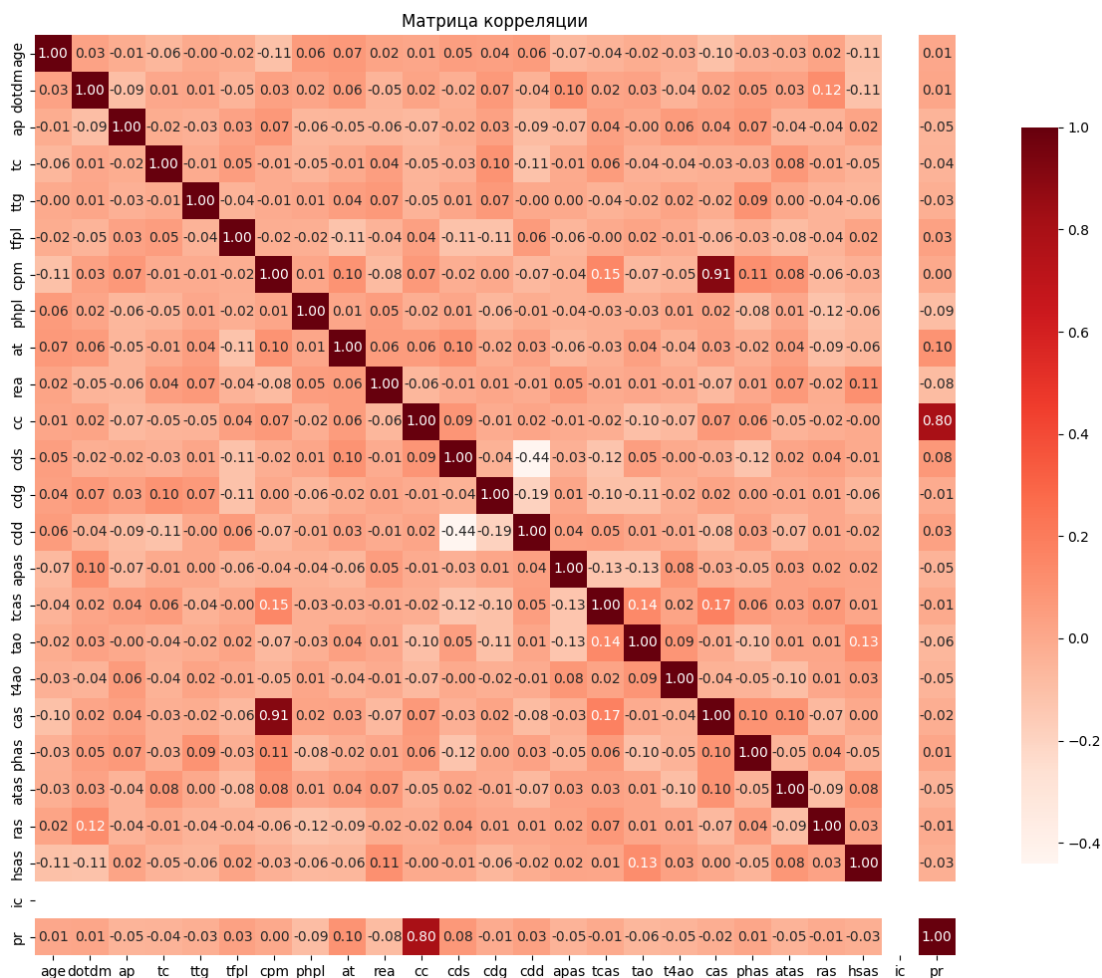


Рис. 3. Матрица корреляции признаков

Усы диаграммы показывают диапазон значений, находящихся в пределах $1,5 \cdot IQR$ от квартилей. Значения, выходящие за эти границы, обозначаются как выбросы и отображаются отдельными точками.

Соответствующая диаграмма изображена на рис. 4. Перед построением диаграммы данные были нормированы. По диаграмме на рис. 4 видно, что выбросы в наших данных присутствуют в признаках: ТТГ (по диаграмме – «ttg»), цитологическая классификация после ТАБ по системе Bethesda (диагностическая категория от 1 до 5) («cc») и кальций общий после операции («tcas»).

Для удаления выбросов можно использовать несколько простых методов. Один из них – межквартильный размах (IQR). Сначала вычисляются границы на основе первых и третьих квартилей, а затем удаляются значения, выходящие за эти границы. Другой подход – стандартное отклонение: если значение слишком далеко от среднего, его можно исключить.

Был выбран метод IQR вместо стандартного отклонения, потому что он более устойчив к влиянию выбросов. Стандартное отклонение может быть искажено экстремальными значениями, что приводит к неправильному

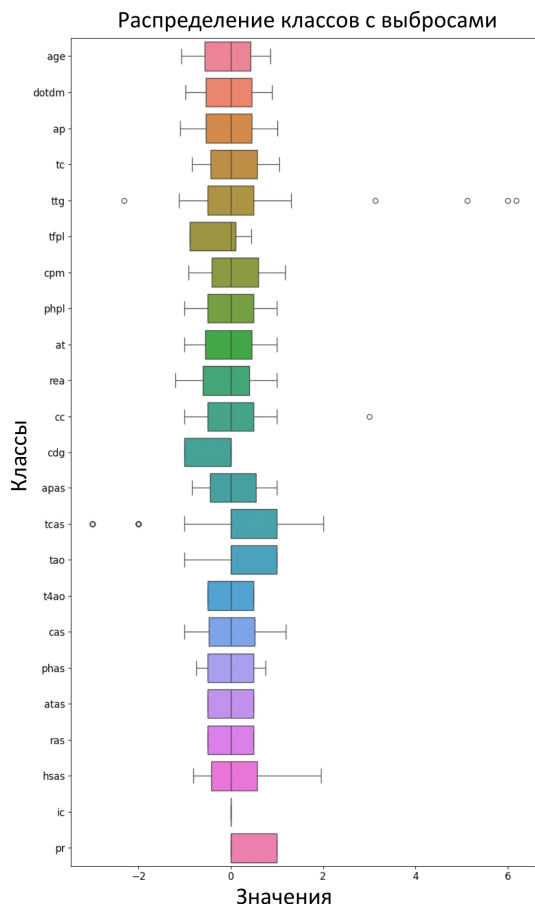


Рис. 4. Распределение классов

определению границ для удаления выбросов. В отличие от этого, IQR фокусируется на центральной части данных и позволяет выявлять аномалии, не завися от крайних значений. Это делает IQR более надежным инструментом для очистки данных и улучшения качества модели.

Описание метода межквартильного размаха (IQR)

1. Определение квартилей: сначала вычисляются первый (Q_1) и третий (Q_3) квартили. Первый квартиль – это значение, ниже которого находится 25 % данных, а третий квартиль – 75 %.

2. Расчет IQR: межквартильный размах (IQR) определяется как разница между Q_3 и Q_1 :

$$IQR = Q_3 - Q_1.$$

3. Выявление выбросов: Значения, которые находятся ниже $Q_1 - 1.5 \cdot IQR$ или выше $Q_3 + 1.5 \cdot IQR$, считаются выбросами и удаляются из набора данных. Таким образом, из 300 изначальных строк в таблице осталось только 272.

С целью повышения точности и надежности предиктивной модели был проведен следующим способом поиск дублированных образцов:

1. Подсчет дубликатов: сначала определяется, сколько раз каждая уникальная строка встречается в наборе данных.

2. Фильтрация результатов: остаются только те строки, которые встречаются более одного раза, то есть только дубликаты.

3. Формирование вывода: если дубликаты найдены, то создается текстовое описание для каждой группы дубликатов, указывая количество повторений и характеристики этих строк. Если дубликатов нет, выводится сообщение о том, что их не обнаружено.

Для рассматриваемого набора данных дубликаты были не найдены.

Далее был произведен поиск и удаление столбцов, заполненных одним значением. В наборе данных были найдены и удалены столбцы (способом далее), значения которых для всех объектов полностью совпадают. Такие столбцы не содержат вариативности, а значит, не несут полезной информации для анализа или обучения модели. Их присутствие может создавать избыточность в данных и не влияет на качество предсказаний, поэтому они исключены для оптимизации работы модели.

Поиск осуществлялся следующим образом:

1. Поиск столбцов: определяются столбцы в наборе данных, в которых все значения одинаковы, то есть количество уникальных значений равно 1. Они заносятся в список.

2. Вывод результатов: затем осуществляется проход по найденным столбцам и вы-

водится информация о каждом из них, указывая, что столбец заполнен одним значением и какое это значение.

Ими оказались признаки: диагноз по TNM. M (после переименования это «dbtm»), сопутствующие заболевания («cd»), диагноз по TNM после гистологии. N («dtahn»), диагноз по TNM после гистологии. M («dtahm») и интраоперационные осложнения («ic»).

Удаление найденных столбцов произведено путем перезаписи исходного набора данных на тот же набор с исключенными столбцами по списку, который был составлен при поиске. В нашем случае их оказалось 5, и из 33 столбцов осталось 28.

После обучения модели была проведена оценка ее производительности, используя метрики качества модели, такие как точность (accuracy), полнота (recall) и точность (precision). Они вычисляются на основе результатов классификации, представленных в виде матрицы ошибок, которая включает четыре категории:

- истинно положительные (TP): количество правильно предсказанных положительных случаев;
- ложно положительные (FP): количество неправильно предсказанных положительных случаев;
- истинно отрицательные (TN): количество правильно предсказанных отрицательных случаев;
- ложно отрицательные (FN): количество неправильно предсказанных отрицательных случаев.

Точность (Accuracy) измеряет, насколько часто модель дает правильные прогнозы. Она рассчитывается по формуле

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Это значение показывает долю всех правильных предсказаний относительно общего числа прогнозов.

Полнота (Recall) отражает способность модели находить все положительные случаи. Она вычисляется следующим образом:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Эта метрика показывает, какую долю всех фактических положительных случаев модель смогла правильно идентифицировать.

Точность (Precision) измеряет долю правильно предсказанных положительных случаев среди всех случаев, которые модель классифицировала как положительные. Она рассчитывается по формуле

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Эти метрики позволяют понять, насколько хорошо модель справляется с задачей, а также выявить возможные проблемы, такие как переобучение или недостаточная способность к выявлению положительных классов. Полученные расчеты представлены в табл. 3. Кроме того, был создан классификационный отчет, который предоставляет собой более детальную информацию о производительности модели для каждого класса. Он представлен в табл. 4.

Далее была составлена матрица ошибок. Она позволяет оценить, как модель классифицирует данные, показывая распределение истинных и предсказанных значений. Она включает правильные и ошибочные предсказания для каждого класса. Данная матрица представлена на рис. 5.

Из полученных результатов видно, что обученная модель демонстрирует высокую точность целевого признака. Доля пациентов с послеоперационным рецидивом, правильно идентифицированных моделью, составила 98 % от общего числа пациентов с рецидивом, а доля пациентов без рецидива, верно классифицированных моделью «как пациенты,

Таблица 3

Метрики качества модели

Метрика	Значение
Accuracy	0,963
Recall	0,963
Precision	0,964

Таблица 4

Классификационный отчет

Precision	Recall	F1-score
0,957	0,978	0,967
0,972	0,946	0,959
0,963	0,963	0,963
0,964	0,962	0,963
0,964	0,963	0,963

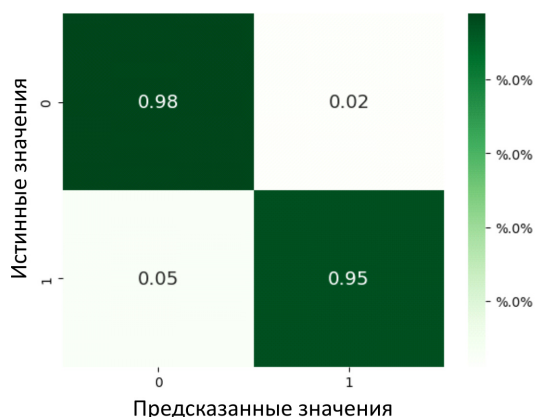


Рис. 5. Матрица ошибок

не имеющие риска рецидива», – 95 % от всех пациентов без рецидива. Это свидетельствует, что модель эффективно справляется с задачей классификации на основе медицинских параметров, что может быть особенно важно для принятия решений в клинической практике. Высокая точность указывает на надежность модели и ее способность правильно идентифицировать случаи рецидива, что может способствовать улучшению диагностики и лечения.

Выводы

В рамках исследования была разработана модель машинного обучения для предсказания высокой вероятности рецидива рака щитовидной железы на основе анализа медицинских параметров⁴. Процесс разработки начался с тщательной предобработки данных, что является критически важным этапом в построении надежных моделей. В ходе предобработки были удалены выбросы и столбцы, содержащие однообразные значения, что позволило улучшить качество данных и избежать искажений в обучении модели. Также была проведена кодировка категориальных переменных, что обеспечило возможность их корректного использования в алгоритмах машинного обучения, и исключены коррелирующие признаки, чтобы минимизировать мультиколлинеарность и повысить интерпретируемость модели.

Для выбора наиболее подходящей модели был проведен сравнительный анализ нескольких алгоритмов классификации. В результате был выбран метод случайного леса, который продемонстрировал высокую эффективность в решении задачи классификации. С использованием метода случайного поиска гиперпараметров была осуществлена оптимизация модели, это позволило определить наилучшие параметры для повышения ее производительности. Полученная модель достигла точности предсказаний 96 %, что свидетельствует о ее высокой надежности и способности к правильной классификации.

⁴ Полиданов М.А., Петрунькин Р.П., Кудашкин В.Н., Волков К.А., Кравченя А.Р., Рафеева П.Д., Трухина М.К., Капралов С.В., Амиров Э.В., Масляков В.В. Система для прогнозирования возникновения рецидивов после оперативного вмешательства при раке щитовидной железы: свидетельство о регистрации программы для ЭВМ № 2024689824 от 11.12.2024. Заявка от 28.11.2024.

Результаты исследования подчеркивают потенциал применения машинного обучения в области медицины, особенно в контексте ранней диагностики и мониторинга за-

болеваний. Высокая точность модели может значительно улучшить процесс принятия клинических решений и повысить качество медицинской помощи.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК / REFERENCES

1. Берштейн Л.М. Рак щитовидной железы: эпидемиология, эндокринология, факторы и механизмы канцерогенеза. Практическая онкология 2007; 8 (1): 1–8 / *Berstein L.M.* Thyroid cancer: epidemiology, endocrinology, factors and mechanisms of carcinogenesis. *Praktical Onkology* 2007; 8 (1): 1–8 (in Russian).
2. Лушников Е.Ф., Цыб А.Ф., Ямасита С. Рак щитовидной железы в России после Чернобыля. М.: Медицина, 2006; 128. / *Lushnikov E.F., Tsyb A.F., Yamashita S.* Thyroid cancer in Russia after Chernobyl. Moscow: Medicine 2006; 128 (in Russian).
3. Bentz B.G. et al. B-RAF V600E mutational analysis of fine needle aspirates correlates with diagnosis of thyroid nodules. *Otolaryngol. Head Neck Surg.* 2009; 140 (5): 709–714.
4. Барчук А.С. Рецидивы дифференцированного рака щитовидной железы. Практическая онкология. 2007; 8 (1): 35. / *Barchuk A.S.* Recurrences of differentiated thyroid cancer. *Practical Oncology* 2007; 8 (1): 35 (in Russian).
5. Amin M.B., Greene F.L., Edge S.B. et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more «personalized» approach to cancer staging. *CA Cancer J Clin.* 2017; 67 (2): 93–99.
6. Kane S.M., Mulbern M.S., Pourshabidi L.K. et al. Micronutrients, iodine status and concentrations of thyroid hormones: a systematic review. *Nutr Rev.* 2018; 76 (6): 418–431.
7. Agretti P. et al. MicroRNA expression profile helps to distinguish benign nodules from papillary thyroid carcinomas starting from cells of fine-needle aspiration. *J. Eur. Endocrinol.* 2012; 167 (3): 393–400.
8. Румянцев П.О., Ильин А.А., Румянцева У.В. и др. Рак щитовидной железы: современные подходы к диагностике и лечению. М.: ГЭОТАР-Медиа 2009; 448. / *Rumyantsev P.O., Ilyin A.A., Rumyantseva U.V.* et al. Thyroid cancer: modern approaches to diagnosis and treatment. Moscow: GEOTAR-Media 2009; 448 (in Russian).
9. Bellevicine C. et al. Cytological and molecular features of papillary thyroid carcinoma with prominent hobnail features: a case report. *Acta Cytol.* 2012; 56 (5): 560–564.
10. Elisei R. et al. The BRAFV600E mutation is an independent, poor prognostic factor for the outcome of patients with low-risk intrathyroid papillary thyroid carcinoma: single-institution results from a large cohort study. *J. Clin. Endocrinol. Metab.* 2012; 97 (12): 4390–4398.
11. Макарыш В.А. Рак щитовидной железы: пособие для пациентов. М. 2016; 168. / *Makarin V.A.* Thyroid cancer. A manual for patients. Moscow 2016; 168 (in Russian).
12. Клинические рекомендации. Дифференцированный рак щитовидной железы. Кодирование по Международной статистической классификации болезней и проблем, связанных со здоровьем: С 73. Возрастная группа: взрослые. М. 2020. / *Clinical guidelines. Differentiated thyroid cancer. Coding according to the International Statistical Classification of Diseases and Related Health Problems: C 73. Age group: adults.* Moscow 2020 (in Russian).

Финансирование. Исследование не имело спонсорской поддержки.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Вклад авторов:

Полиданов М.А., Петрунькин Р.П., Кудашкин В.Н., Кравченя А.Р., Масляков В.В., Капралов С.В. – концепция и дизайн исследования.

Полиданов М.А., Петрунькин Р.П., Кудашкин В.Н., Волков К.А., Кравченя А.Р., Масляков В.В., Капралов С.В., Асланов Г.Э., Лосякова Е.В., Обухов И.С., Осина А.Д., Курмаева А.К. – сбор и обработка материала.

Барулина М.А., Бендик И.Ю., Коваленко И.И., Полиданов М.А. – статистическая обработка.

Барулина М.А., Бендик И.Ю., Коваленко И.И., Полиданов М.А., Петрунькин Р.П., Кудашкин В.Н., Волков К.А., Кравченя А.Р., Масляков В.В., Капралов С.В., Асланов Г.Э., Лосякова Е.В., Обухов И.С., Осина А.Д., Курмаева А.К. – редактирование.

Все авторы утвердили окончательный вариант текста статьи.

Ограничение исследования. Разрешение на проведение исследования получено от локального этического комитета (ЛЭК) Медицинского университета «Реавиз» (протокол № 9 от 10.09.2024). Исследование проводили при наличии добровольных информированных согласий пациентов в соответствии с декларацией о соблюдении международных, а также российских этических принципов и норм (выписка из протокола № 19 заседания комитета по биоэтике от 26 октября 2018 г.). Исследование выполнено в соответствии с требованиями Хельсинкской декларации Всемирной медицинской ассоциации (в ред. 2013 г.).

Поступила: 05.03.2025

Одобрена: 08.05.2025

Принята к публикации: 23.05.2025

Просьба ссылаться на эту статью в русскоязычных источниках следующим образом: Возможность предсказания вероятности рецидива рака щитовидной железы методами машинного обучения / М.А. Барулина, И.Ю. Бендик, И.И. Коваленко, М.А. Полиданов, Р.П. Петрунькин, В.Н. Кудашкин, К.А. Волков, А.Р. Кравченя, В.В. Масляков, С.В. Капралов, Г.Э. Асланов, Е.В. Лосякова, И.С. Обухов, А.Д. Осина, А.К. Курмаева // Пермский медицинский журнал. – 2025. – Т. 42, № 3. – С. 130–143. DOI: 10.17816/pmj423130-143

Please cite this article in English as: Barulina M.A., Bendik I.Yu., Kovalenko I.I., Polidanov M.A., Petrunkin R.P., Kudashkin V.N., Volkov K.A., Kravchenya A.R., Maslyakov V.V., Kapralov S.V., Aslanov H.E., Losyakova Ye.V., Obukhov I.S., Osina A.D., Kurmaeva A.K. Possibility of predicting the probability of thyroid cancer recurrence by machine learning methods. *Perm Medical Journal*, 2025, vol. 42, no. 3, pp. 130-143. DOI: 10.17816/pmj423130-143