DOI: 10.12731/2658-6649-2023-15-3-475-496
УДК 004.85:61.616

Научная статья | Медицинская информатика

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЗАБОЛЕВАНИЙ СЕРДЦА

А.И. Павлова

Работа посвящена применению алгоритмов машинного обучения для прогнозирования сердечно-сосудистых заболеваний (ССЗ). Ежегодно во всем мире фиксируется большое количество смертей. По данным Всемирной организации здравоохранения ССЗ являются основной причиной высокой смертности в мире. Одним из необходимых профилактических мер по снижению смертности от ССЗ является своевременное прогнозирование заболеваний у людей, подвергшихся высокому риску таких заболеваний.

Обоснование. *Для своевременного прогнозирования ССЗ в настоящее время используют специально разрабатываемые шкалы и алгоритмы машинного обучения. Для прогнозирования заболеваний сердца часто применяют алгоритмы: наивный байесовский классификатор (Naïve Bayes Classificator, NBC), k-ближайших соседей (K-Nearest Neghboors, KNN), дерево решений (Decision Tree, DT). В отечественной литературе известны работы, посвященные применению прогнозированию ССЗ с помощью градиентного алгоритма Adam при обучении глубокой нейронной сети. Одним из необходимых условий повышения прогностической способности модели машинного обучения (ММО) является оптимальный подбор гиперпараметров. Выбор оптимальных гиперпараметров ММО часто осуществляется на основании эмпирического опыта.*

Цель. *Изучить особенности применения машинных алгоритмов для прогнозирования заболеваний сердца.*

Материалы и методы. *Научная новизна работы. В проведенных исследованиях выполнен анализ алгоритмов машинного обучения для прогнозирования риска возникновения ССЗ с применением подхода автоматического поиска гиперпараметров ММО. Для построения ММО использованы следующие алгоритмы: NBC, KNN, DT, логистическая регрессия (Logistic Regression), машина опорных векторов (Support Vector Machine, SVM), случайные леса (Random Foorest, RF), адаптированный полиномиальный байесовский классификатор*

(Complement Naïve Bayes Classificator, CNBC), линейный дискриминантный анализ (Linear Discriminant Analysis, LDA), градиентный бустинг (XGBoost).

Для оценки точности моделей машинного обучения использованы показатели: средняя абсолютная ошибка (mean absolute error, MAE), точность (precision), полнота (recall), F-мера, доля ложноположительных примеров (False Positive Rate, FPR), доля отрицательных примеров (False Negative Rate, FNR). Дополнительно при анализе результатов построения ММО служил визуальный анализ кривой ROC (receiver operating characteristic) и площадь под кривой ROC (Areas under the curve, AUC). Использование значения AUC позволяет оценить прогностические возможности ММО.

Результаты. *Результаты обучения показали, что алгоритмы RF и XGBoost характеризуются более высокими показателями точности. При оптимальном подборе параметров ММО общая точность классификации составила 0,88 и 0,94 соответственно.*

Заключение. *Применение алгоритмов машинного обучения позволяет с высокой точностью построить прогнозные модели. Ансамблевые алгоритмы машинного обучения RF и XGBoost характеризуются более высокими показателями точности в сравнении со следующими алгоритмами: деревья решений, байесовские методы классификации, логистическая регрессия, линейный дискриминантный анализ.*

Ключевые слова: *сердечно-сосудистые заболевания; алгоритмы машинного обучения; модель машинного обучения; прогнозирование*

Для цитирования. *Павлова А.И. Применение алгоритмов машинного обучения для прогнозирования заболеваний сердца // Siberian Journal of Life Sciences and Agriculture. 2023. Т. 15, №3. С. 475-496. DOI: 10.12731/2658-6649-2023-15-3-475-496*

Original article | Medical Informatics

APPLICATION OF MACHINE LEARNING ALGORITMS FOR HEART DISEASE PREDICTION

A.I. Pavlova

This paper focuses on the application of machine learning algorithms to predict cardiovascular diseases (CVDs). Every year a large number of deaths are registered all over the world. According to the World Health Organisation, CVDs are the leading cause of high mortality in the world. One of the necessary preventive measures

to reduce mortality from CVDs is the timely prediction of diseases in people at high risk of such diseases. Specially developed scales and machine learning algorithms are now being used for the timely prediction of CVDs.

Background. *To predict heart disease, algorithms are often used: naive Bayesian classifier (Gaussian Naïve Bayes Classificator, GNBC), k-nearest neighbours (K-Nearest Neghboors, KNN), decision tree (Decision Tree, DT). In domestic literature, there are known works devoted to the application of SWD prediction using Adam gradient algorithm in deep neural network training. One of the necessary conditions for increasing the predictive ability of a machine learning model (MLM) is the optimal selection of hyperparameters. The choice of the optimal hyperparameters is often made on the basis of empirical experience.*

Purpose. *To explore the specific application of machine algorithms to the prediction of heart disease.*

Materials and methods. *Scientific novelty of the work. In this research we analyse machine learning algorithms for predicting the risk of CVDs using the approach of automatic search for hyperparameters MMO. The following algorithms were used to construct MMOs: NBS, KNN, DT, Logistic Regression, Support Vector Machine (SVM), Random Foorest (RF), Complement Naïve Bayes Classificator (CNBC), Linear Discriminant Analysis (LDA), Radial Basic Function (RBF), Gradient Boost (XGBoost). To evaluate the accuracy of machine learning models we used the following indicators: mean absolute error (MAE), precision, completeness (recall), F-measure (F-beta), False Positive Rate (FPR), False Negative Rate (FNR). Additionally, visual analysis of ROC curve (receiver operating characteristic) and areas under the curve (areas under the curve, AUC) were used to analyse the results of MMO. Using AUC value allows to estimate prognostic ability of MLM.*

Results. *The training results showed that RF and XGBoost algorithms are characterized by higher accuracy. With optimal selection of MMO parameters, the overall classification accuracy was 0.88 and 0.94 respectively.*

Conclusion. *The application of machine learning algorithms allows predictive models to be built with high accuracy. This requires the construction of a machine learning model. The ensemble machine learning algorithms RF and XGBoost have higher accuracy rates than the following algorithms: decision trees, Bayesian classification methods, logistic regression, linear discriminant analysis.*

Keywords: *cardiovascular diseases; machine learning algorithms; machine learning model; prediction*

For citation. *Pavlova A.I. Application of Machine Learning Algorithms for Heart Disease Prediction. Siberian Journal of Life Sciences and Agriculture, 2023, vol. 15, no. 3, pp. 475-496. DOI: 10.12731/2658-6649-2023-15-3-475-496*

Введение

Сердечно-сосудистые заболевания (ССЗ) остаются одной из основных причин высокой смертности людей в мире. По данным Всемирной организации здравоохранения ежегодно фиксируют 17,3 млн. смертей. Большую часть (45% от всего количества смертей в год) составляют группы заболеваний: ССЗ, онкологические, бронхолегочные и сахарный диабет. Основными факторами наступления смерти от ССЗ являются сердечный приступ и инсульт [19].

В России ежегодно фиксируется высокая смертность среди взрослого населения на протяжении многих десятилетий. Более 80% от общего количества смертей обусловлено ишемической болезнью сердца и мозговыми инсультами [9, 20]. Артериальная гипертония также является одним из главных факторов развития ССЗ [16, 20]. Профилактика ССЗ направлена на предупреждение и лечение болезней сердца. Развитие и принятие профилактических мер в современном понимании связывают с диагностированием, разработкой корректных шкал для оценки болезней сердца, использованием современных методов прогнозирования ССЗ и др. [2, 11, 12, 14]. Своевременное прогнозирование сердечно-сосудистых заболеваний позволяет более эффективно использовать лечебные ресурсы, включающее оперативное хирургическое лечение и дорогостоящее высокотехнологичное оборудование [9, 11, 13].

В зарубежной литературе для прогнозирования риска возникновения ССЗ применяют алгоритмы машинного обучения [16, 21, 22, 25, 28-30]. Машинное обучение включает различные классификаторы, используемые для контролируемого, неконтролируемого или ансамблевого обучения. За последние несколько лет подходы, включающие машинное обучение, оказывают значительное влияние на обнаружение и диагностику заболеваний в медицине [22, 37, 39]. В работе [34] успешно продемонстрированы возможности машинного обучения при прогнозировании ССЗ байесовскими методами. Авторами работы [35] было выявлено, что метод деревьев решений (Decision Tree, DT) имеет более высокие показатели точности в сравнении с наивным байесовский классификатор (Gaussian Naïve Bayes Classificator, GNBC). В работе [35] выполнен анализ следующих алгоритмов машинного обучения для прогнозирования болезней сердца: GNBC, KNN, метод деревьев решений, метод стохастического градиентного спуска, машина опорных векторов (Support Vector Machine, SVM), адаптивного бустинга AdaBoost. Для улучшения обучающей способности моделей машинного обучения (ММО) автором были использованы различные размеры валидационной выборки. В результате выявлено, что алгоритм KNN обладает наиболее более высокой прогностической спо-

собностью (точность результатов обучения составила 99,71%) в сравнении с другими алгоритмами. Метод KNN часто используется в зарубежной литературе и характеризуется простотой программной реализации [4]. В работах [17, 45] продемонстрировано, что методы SVM и случайный лес (Random Forest, RF) обладают более высокими показателями точности и скорости обучения в отличие от метода KNN.

В отечественной литературе для прогнозирования ССЗ предлагают разрабатывать специальные шкалы, необходимые на этапе оценки признаков машинного обучения, а также использовать глубокие искусственные нейронные сети [2, 10-12, 14].

В целом, подход машинного обучения предполагает «обучение» алгоритма на контрольном наборе данных, для которого известен статус заболевания (заболевание или отсутствие заболевания), а затем применение обученного алгоритма к переменному набору данных для прогнозирования статуса заболевания у пациентов, для которых он еще не определен. Более точное прогнозирование заболевания ССЗ с помощью алгоритмов машинного обучения позволит врачам улучшить обнаружение, диагностику, классификацию, стратификацию риска, потенциально минимизирует необходимое клиническое вмешательство при лечении пациентов.

При машинном обучении актуальны вопросы оптимального выбора гиперпараметров ММО. Под гиперпараметрами понимают параметры алгоритмов машинного обучения, значения которых устанавливаются перед обучением ММО [6]. Гиперпараметры используются в процессе обучения и поэтому служат для управления процессом обучения. Неправильный выбор гиперпараметров часто создает проблемы недостаточной или чрезмерной подгонки ММО. Выбор гиперпараметров можно осуществлять эмпирическим путем. Однако для более эффективной настройки ММО требуется большое количество валиадационных данных, ранее не участвовавших в обучении. Ручной способ выбора гиперпараметров ММО требует существенных трудовых и временных затрат. Поэтому актуальны работы, посвященные изучению выбора гиперпараметров ММО в автоматическом режиме.

Машинное обучение без учителя часто применяют для снижения размерности входного признакового пространства. Алгоритмы машинного обучения без учителя предусматривают кластеризацию данных с разбиением их на однородные группы (классы, кластеры). Особенность применения алгоритмов машинного обучения состоит в отсутствии меток классов, позволяющих соотнести обучающие примеры к заранее известному классу. Разделение данных на классы производится на основании разнообразных

метрик близости. Наиболее известной метрикой близости в статистическом анализе данных и машинном обучении принято Евклидово расстояние [1, 5]. Отсутствие требований формирования обучающих выборок с метками классов делает привлекательными алгоритмы без учителя для решения задач кластеризации, снижения размерности данных, обнаружения выбросов, поиске закономерности распределения входных данных [8].

Машинное обучение с учителем. Алгоритмы машинного обучения с учителем предусматривают контролируемое обучение с использованием обучающего набора данных (обучающей выборки). Набор данных включает конечное множество объектов (примеров). Для решения задачи классификации алгоритмами машинного обучения должен быть задан определенный класс объектов: в обучающей выборке каждый пример описывается признаками и меткой. Метка используется для оценки принадлежности примеров к определенному классу объектов и может быть задана в виде целочисленных ответов 0 или 1 (0 – соответствует ответу «обучающий пример не принадлежит к определенному классу» и 1 – соответствует ответу «обучающий пример принадлежит к классу»).

При прогнозировании риска возникновения ССЗ алгоритмы машинного обучения используются для бинарной классификации. Входные признаки часто имеют числовой или категориальный характер.

Новизна исследований состоит в применении алгоритмов машинного обучения для прогнозирования сердечно-сосудистых заболеваний с применением подхода автоматического поиска гиперпараметров моделей машинного обучения.

Материалы и методы исследований

Для построения модели машинного обучения использовали набор данных с 1025 примерами, представляющих собой описание состояния 1025 пациентов по 13 признакам (таблица 1), влияющим на риск возникновения ССЗ. Данные признаки описывают состояние пациента наличие типа боли в грудной клетке, уровень артериального давления и др. и используются в качестве предикторов в алгоритмах машинного обучения. Процедура прогнозирования ССЗ у пациента осуществлялась алгоритмами машинного обучения с учителем, для вычислительной работы которых каждому примеру было определено выходное значение целевой переменной. Значение целевой переменной target представляет собой метку классов и принимает два значения: 0 или 1, соответствующее отсутствию сердечно-сосудистых заболеваний или ее наличию соответственно.

Таблица 1.

Признаки, использованные для характеристики состояния пациента

Признак	Обозначение	Единицы измерения
Возраст (полных лет)	age	год
Пол пациента	sex	1 – мужской, 0 – женский
Тип боли в груди	cp	0 – типичная стенокардия 1 – атипичная стенокардия 2 – неангинальная боль 3 – бессимптомная
Уровень артериального давления	trestbps	мм рт.ст. миллиметры ртутного столба
Холестерин в сыворотке крови	chol	мг/дл
Сахар в крови	fbs	1 – более 120 мг/дл; 0 – менее более 120 мг/дл
Результаты электрокардиограммы в состоянии покоя	restecg	0 – норма, 1 – наличие аномалии ST-T волн, инверсии T волн и/или подъем или депрессия ST более 0,05 мВ, 2 – наличие вероятной или определенной гипертрофии левого желудочка согласно критериям Estes (Эстеса)
Максимальная частота сердечных сокращений	thalach	удары в минуту
Обнаружение стенокардии, вызванной физической нагрузкой	exang	1 – присутствует, 0 – отсутствует
Обнаружение у пациента депрессии сегмента ST, вызванной физической нагрузкой по сравнению с покоем	oldpeak	1 – присутствует, 0 – отсутствует
Максимальная нагрузка ST	slope	максимальная нагрузка ST Вверх: наклон вверх, плоский: плоский, вниз: наклон вниз
Количество крупных сосудов, окрашенных при флоуроскопии	ca	1 – от 0 до 3; 0 – отсутствуют
Наличие дефекта	thal	0 – норма; 1 – фиксированный дефект; 2 – обратимый дефект
Наличие ССЗ	target	1 – присутствует, 0 – отсутствует

В работе использованы алгоритмы машинного обучения с учителем:

- наивный байесовский классификатор (Gaussian Naïve Bayes Classificator, GNBC);
- адаптированный полиномиальный байесовский классификатор (Complement Naïve Bayes Classificator, CNBC);
- линейный дискриминантный анализ (Linear Discriminant Analysis, LDA);
- k-ближайших соседей (K-Nearest Neghboors, KNN) [25];
- деревья решений (Decision Tree, DT) [32,];
- логистическая регрессия (Logistic Regression) [3,4,8];
- машина опорных векторов (Support Vector Machine, SVM) [19,21,36,38,40,44];
- случайный лес (Random Foorest, RF) [18,27,28,43];
- градиентного бустинга деревьев решений (XGBoost).

Процесс прогнозирования риска возникновения ССЗ методами машинного обучения включал следующие этапы:

1) импорт библиотек машинного обучения Numpy, Pandas, Scikit-learn, Matplotlib and Seaborn;
2) предварительная подготовка данных для обучения, связана с поиском и заполнением пропусков в данных;
3) нормализация входных данных с использованием функции StandardScalar библиотеки Scikit-learn;
4) разделение набора входных данных на обучающую (80% от общего количества примеров) и тестовую выборки (20% от общего количества примеров);
5) обучение модели машинного обучения с параметрами, заданными в библиотеке Scikit-learn по умолчанию;
6) подбор оптимальных параметров модели машинного обучения;
7) тестирование результатов машинного обучения на тестовой выборке данных;
8) оценка точности результатов машинного обучения.

Предварительная подготовка данных для обучения включала анализ распределения данных и предусматривала: анализ пропусков в данных, заполнение пропусков в случае из наличия, визуальный анализ распределения данных, корреляционный анализ данных.

Процесс подготовки данных для машинного обучения включал разделение данных на три выборки, используемых для различных целей: для обучения, тестирования и валиадации. Для обучения было использовано

70% (718 примеров), для тестирования 15% (307 примеров) и для валидационной оценки 15% (307 примеров) от общего количества примеров.

Оценка качества модели машинного обучения осуществлялась по следующим параметрам: accuracy, precision, F1-Score, macro-averaging, ROC и AUC [3, 6, 7, 15].

Для оценки качества ММО применялся визуальный и количественных анализ ROC-кривой (Receiver Operator Characteristic), отражающих специфичность и чувствительность ММО. График ROC-кривой использовали в теории обнаружения сигналов при определения зависимости частоты правильно распознанных сигналов от частоты неправильно распознанных сигналов [24, 41]. ROC-кривая представляет собой функцию частоты истинно положительных результатов (чувствительность) от частоты ложноположительных результатов (специфичность). Отдельная точка на данном графике отображает возможности ММО в виде пары чувствительность/специфичность, соответствующую принятому порогу [3, 13, 46]. При визуальной оценке ROC-кривой принято считать, чем выше и левее она расположена относительно осей координат, тем лучшими прогностическими свойствами характеризуется ММО. ROC-кривая применятся при тестировании различных заболеваний в медицине [33, 42].

Для количественного анализа вычислялась площадь Area Under Curve (AUC), ограниченная справа и снизу осями координат, слева полученными точками классифицированных примеров. AUC определена с помощью библиотеки Scikit-learn методом трапеций. Для оценки значений AUC служила экспертная шкала (табл.2).

Таблица 2.

Шкала для оценки значений AUC [46]

Интервал	Качество ММО
0,5-0,6	Неудовлетворительное
0,6-0,7	Среднее
0,7-0,8	Хорошее
0,8-0,9	Очень хорошее
0,9-1,0	Отличное

Результаты исследований

Построены графики распределения частот входных признаков машинного обучения и выходной переменной (рис. 1). Распределение исходных данных не подчиняется типичным формам распределения – равномерному и симметричному.

По результатам применения инженерной библиотеки Pandas Python было выявлено, что в исходном наборе данных отсутствуют пропуски (рис. 1). Проверка пропусков производилась с помощью метода isnull, позволяющая вычислить логическое значение в виде двумерной матрицы с ответами True или False, а также в виде одномерной матрицы (рис. 2).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1021	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1022	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1023	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1024	False	False	False	False	False	False	False	False	False	False	False	False	False	False

Рис. 1. Проверка пропущенных значений в наборе данных

На рис. 2 представлено количество пропущенных значений по столбцам в наборе данных, вычисленных с помощью Pandas.

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Рис. 2. Результаты проверки пропусков в наборе данных

Распределение обучающих примеров по двум классам объектов, соответствующих положительным ответам составило 526, а отрицательными

ответами 499 (рис.3), поэтому разделение примеров можно считать сбалансированным.

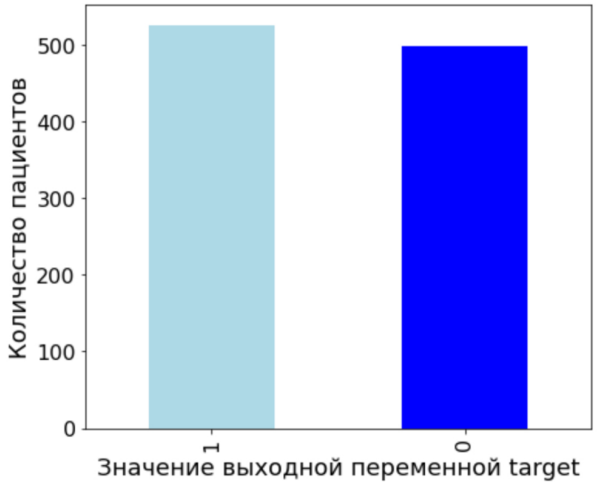

Рис. 3. Гистограмма распределения выходной переменной по двум классам объектов (1 – положительный ответ, соответствующий наличию ССЗ у пациента; 0 – отрицательный ответ, соответствующий отсутствию ССЗ у пациента)

Наличие ССЗ обнаруживается в большинстве случаев у мужчин (на 100 случаев больше чем у женщин) (рис. 4).

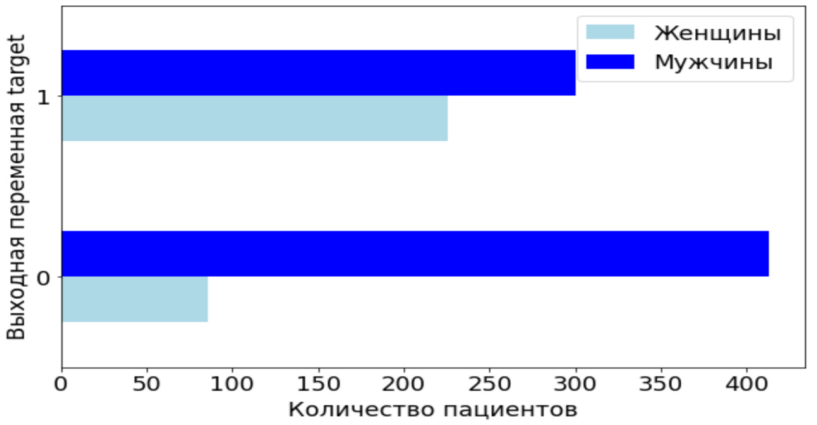

Рис. 4. Количество пациентов с ССЗ среди мужчин и женщин

Результаты корреляционного анализа показали, что линейная зависимость между входными переменными и выходной переменной слабая (рис. 5). Наиболее выражена положительная корреляционная зависимость между типом боли в грудной клетке (0,43), максимальной частотой сердечных сокращений (0,42), максимальной нагрузкой ST (0,35).

Переменные, описывающие наличие стенокардии, вызванной физической нагрузкой и депрессии сегмента ST оказываются отрицательно коррелированы с выходной переменной с коэффициентом корреляции равным -0,44. При этом переменные, описывающие обнаружение у пациента более 3 крупных сосудов, окрашенных при проведении флоуроскопии и наличие дефекта характеризуются слабой линейной отрицательной связью с риском возникновения ССЗ с коэффициентами, равными 0,38 и 0,34 соответственно.

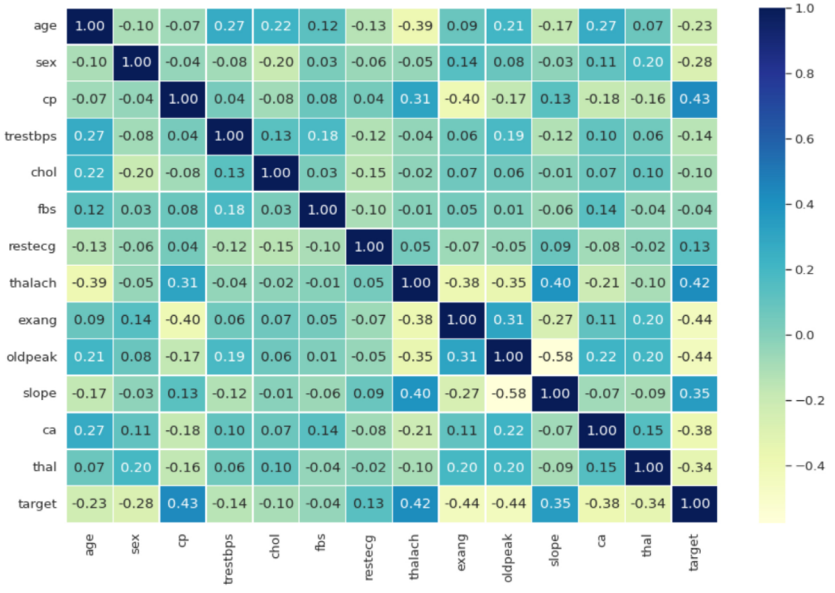

Рис. 5. Корреляционная матрица входных признаков

Для подбора оптимальных гиперпараметров ММО в работе использован метод Grid SearchCV (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html). Процесс подбора параметров производился по «сетке» в переделах определенного диапазона с опреде-

ленной длиной шага. В дальнейшем вычисленные параметры применены для обучения и тестирования модели машинного обучения.

Поиск по «сетке» гиперпараметров с помощью библиотеки Sklearn позволяет принимает на вход модель и различные значения гиперпараметров (сетку гиперпараметров). Для каждого возможного сочетания значений гиперпараметров с помощью метода Grid SearchCV вычисляется ошибка и подбирается сочетание гиперпараметров, при которых значение ошибки будет минимальным. В таблице 3 приведены результаты оценки точности ММО, вычисленные для тестовой и валидационной выборок данных. В таблице 3 приведены усредненные значения показателей точности по двум классам объектов.

ММО для алгоритмов NBC и CNBC имеют наиболее низкие значения точности. Метрики точности данных алгоритмов примерно равны. Средняя точность классификаторов составила 0,71 и 0,72 соответственно. Площадь под кривой ROC равна 0,69 и 0,72 соответственно и согласно экспертной оценке (табл.2) качество ММО можно условно оценить как «среднее».

Метрики точности ММО, рассчитанные для алгоритмов машинного обучения LDA, LG и KNN имеют близкие значения. Средняя точность классификаторов немного больше (на 0,02) в сравнении с байесовскими классификаторами. Площадь под кривой ROC для алгоритмов LDA, LG и KNN также имеет более высокие показатели в сравнении с байесовскими алгоритмами. Согласно табл.2 качество ММО можно охарактеризовать как «хорошее».

ММО, построенные с применением алгоритмов DT и SVM характеризуются хорошими показателями. Вычисленная точность классификаторов составила 0,84 и 0,85 соответственно. Значения F1-меры, полноты, площади под кривой колеблется от 0,74 до 0,86. Средняя ошибка классификаторов равна 0,83 и 0,84 соответственно. Согласно экспертной оценке (табл.2) качество ММО можно условно оценить как «очень хорошее».

Метрики точности ММО для алгоритмов RF и XGBoost имеют наиболее высокие значения. Вычисленная точность классификаторов равна 0,91 и 0,93 соответственно, значение площади под кривой ROC составило 0,90 и 0,96. Согласно вычисленным показателям качество ММО данных алгоритмов можно охарактеризовать как «отличное».

Таблица 3.

Результаты обучения машинного обучения

Алгоритм	Precision	F1-score	Recall	AUC	Mean
LDA	0,74	0,78	0,69	0,72	0,73
NBC	0,68	0,76	0,70	0,69	0,71

Окончние табл. 3.

KNN	0,76	0,67	0,74	0,79	0,74
CNBC	0,69	0,75	0,72	0,72	0,72
LG	0,76	0,77	0,80	0,74	0,77
DT	0,84	0,86	0,82	0,81	0,83
SVM	0,85	0,81	0,83	0,85	0,84
RF	0,91	0,84	0,87	0,90	0,88
XGBoost	0,93	0,95	0,92	0,96	0,94

Примечание: Precision – точность; F1-score – мера F1; Recall – полнота; AUC (Area Under Curve) – площадь под кривой ROC; Mean – средняя ошибка. Алгоритмы машинного обучения: LDA – линейный дискриминантный анализ; NBC – наивный байесовский классификатор; KNN - k-ближайших соседей; CNBC – адаптированный полиномиальный байесовский классификатор; LG – логистическая регрессия; DT – деревья решений; SVM – машина опорных векторов; RF – случайный лес; XGBoost – градиентный бустинг.

Заключение

Высокое значение recall указывает на меньшую склонность к ложноотрицательным результатам. Наиболее низкие значения точности были получены при использовании линейного дискриминантного анализа. Для данной ММО характерны наиболее высокие значения ложных ответов в среднем выше на 5-10% в сравнении с RF и XGBoost.

Полученные результаты вычислений показали невысокую точность для ММО, построенных с помощью байесовских методов классификации. Наивный байесовский классификатор отличается незначительно по точности в сравнении с адаптированный полиномиальный байесовский классификатором на 0,02.

Интуитивно понятные ММО построенные с помощью алгоритмов логистической регрессии и KNN характеризуются примерно одинаковыми показателями точности. Хорошие показатели точности вычислены при использовании машины опорных векторов. Однако для данного алгоритма характерно большее количество ложноположительных ответов (примерно на 4-7%), в сравнении с алгоритмами RF и XGBoost.

Наиболее высокие прогностические способности и точностные характеристики присущи ММО, построенных с помощью алгоритмов машинного обучения – случайный лес RF и градиентный бустинг XGBoost. Точность классификаторов выше 0,85. Другие параметры ММО (recall, F1 score, AUC) тоже являются высокими. При этом XGBoost имеет наилучшие результаты по точности.

Использование множества решающих деревьев при работе алгоритма существенно повышает точность прогнозирования риска возникновения ССЗ в сравнении с методом дерева решений.

Информация о конфликте интересов. Автор заявляет об отсутствии конфликта интересов.

Информация о спонсорстве. Работа не имела спонсорской поддержки, автор не получал гонорар за исследование.

Список литературы

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.
2. Белялов Ф.И. Прогнозирование заболеваний с помощью шкал // Комплексные проблемы сердечно-сосудистых заболеваний. 2018. Т.7. №.1. С. 84–93. https://doi.org/10.17802/2306-1278-2018-7-1-84-93
3. Ветров Д.П., Кропотов Д.А. Алгоритмы выбора моделей и построения коллективных решений в задачах классификации, основанные на принципе устойчивости. Москва, URSS, 2006. 112 с.
4. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин). URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
5. Воронцов К.В. Лекции по статистическим (байесовским) алгоритмам классификации. URL: http://www.ccas.ru/voron/download/Bayes.pdf
6. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. М.: 2013. 387 с.
7. Дуда Р., Харт П. Распознавание образов и анализ сцен / Пер. с англ. М.: Мир, 1976. 511 с.
8. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 270 с.
9. Кардиоваскулярная профилактика 2017. Российские национальные рекомендации // Российский кардиологический журнал. 2018. № 6. С. 7-122. https://doi.org/10.15829/1560-4071-2018-6-7-122.
10. Литвин А.А., Калинин А.Л., Тризна Н.М. Использование данных доказательной медицины в клинической практике (сообщение 3 – диагностические исследования) // Проблемы здоровья и экологии. 2008. Т.18. №4. С.12-19.
11. Невзорова В.А., Плехова Н.Г., Присеко Л.Г. и др. Методы машинного обучения в прогнозировании исходов сердечно-сосудистых заболеваний с

артериальной гипертензией (по материалам ЭССЭ-РФ в Приморском крае) // Российский кардиологический журнал. 2020. Т. 25. №3. С. 10–16. https://doi.org/10.15829/1560-4071-2020-3-3751
12. Панев Н.И., Евсеева Н.А., Филимонов С.Н., Коротенко О.Ю., Данилов И.П. Система прогнозирования развития ишемической болезни сердца у шахтёров с антракосиликозом // Медицина труда и промышленная экология. 2021. Т. 61. №.6. С. 365–370. https://doi.org/10.31089/1026-9428-2021-61-6-365-370
13. Самигулин Т.Р., Джурабаев А.Э.У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии, 2021, Т. 6, №1. С.55–62. https://doi.org/10.18413/2518-1092-2021-6-1-0-7
14. Смирнова М.Д., Свирида О.Н., Фофанова Т.В. и др. Алгоритм прогнозирования сердечно-сосудистых осложнений у больных низкого/умеренного риска с использованием классических и новых факторов (по данным) десятилетнего наблюдения // Кардиоваскулярная терапия и профилактика. 2021. Т. 20. №. 6. С. 2799. https://doi.org/10.15829/1728-8800-2021-2799
15. Ту Дж., Гонсалес Р. Принципы распознавания образов / Пер. с англ.; Пол ред. Ю. И. Журавлева. М.: Мир, 1978. 411 с.
16. Чазова И.Е., Ошепкова Е.В. Опыт борьбы с сердечно-сосудистыми заболеваниями в России // Аналитический вестник. 2015. № 44(597). С.4-8.
17. Aravind Akella, Sudheer Akella Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution // Future Science OA, 2021. https://doi.org/10.2144/fsoa-2020-0206
18. Breiman L. Random Forest // Machine Learning, 2001, vol. 45, no. 1, pp. 5-32.
19. Bordes A., Ertekin S., Weston J., Bottou L. Fast Kernel Classifiers with Online and Active Learning // Journal of Machine Learning Research, 2005, no. 6, pp. 1579–1619.
20. Cardiovascular disease. World Health Organization website. 2022. https://www.who.int/ru/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
21. Cortes C., Vapnik V. Support vector networks // Machine Learning, 1995, no. 20, pp. 273–297.
22. Cuocolo R., Perillo T., De Rosa E., Ugga L., Petretta M. Current applications of big data and machine learning in cardiology // Journal Geriatric Cardiology, 2019, vol. 16, no.8, pp.601 – 607.
23. Deo R.C. Machine learning in medicine // Circulation, 2015, vol. 132, no. 20, pp. 1920- 1930.
24. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters, 2006, vol. 27, no 8, pp. 861-874.

25. Forgy E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications // Biometrics, 1965, vol. 21, pp. 768–769.
26. Foster K.R., Koprowski R., Skufca J.D. Machine learning, medical diagnosis, and biomedical engineering research-commentary // Biomedical Engineering Online, 2014, no. 13, article no. 94. https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-94
27. Guyon, B. Boser, Vapnik V. Automatic capacity tuning of very large VC-dimension classifiers. In S. J. Hanson, J. D. Cowan, and C. Lee Giles, editors, Advances in Neural Information Processing Systems, Morgan Kaufmann, San Mateo, CA, 1993, pp. 147–155.
28. Ho T. K. The Random subspace method for construction decision tree forests // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, vol. 20, no. 8, pp. 832–844. https://doi.org/10.1109/34.709601
29. Karimollah Hajian-Tilaki Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation // Caspian Journal of Internal Medicine, 2013, vol. 4, no. 2, pp. 627–635. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/
30. Khan Y, Qamar U, Yousaf N, Khan A. Machine learning techniques for heart disease datasets // Proceedings of the 2019 11th International Conference on Machine Learning and Computing – ICMLC '19, 2019. https://doi.org/10.1145/3318299.3318343
31. Khaled M.A. Prediction of heart disease and classifiers sensitivity analysis // Almustafa BMC Bioinfirmatics, 2020, no. 21, Article number: 278. https://doi.org/10.1186/s12859-020-03626-y
32. Kohavi R., Quinlan J.R. Decision tree discovery. C5.1.3. 1999. https://ai.stanford.edu/~ronnyk/treesHB.pdf
33. Kummar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers // Indian Pediatrics, 2011, vol. 48, no. 7, pp. 277-89.
34. Lu Y., Dendukuri N., Schiller I., Joseph L. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies // Statistics in medicine, 2010, no. 29, pp. 2532-2543.
35. Maheswari S., Pitchai R. Heart disease prediction system using decision tree and naive bayes algorithm // Current Medical Imaging, 2019, no. 15, pp. 712–77. https://doi.org/10.2174/1573405614666180322141259
36. Nefedov A. Support Vector Machines: A Simple Tutorial. 2016. https://svmtutorial.online/download.php?file=SVM_tutorial.pdf
37. Obermeyer Z., Ezekiel J.E. Predicting the future – big data, machine learning, and clinical medicine // The New England Journal of Medicine, 2016, vol. 375, no.13, pp. 1216–1219.

38. Quinlan J. R. Induction of decision trees // Machine Learning, 1986, vol. 1, no. 1, pp. 81-106.
39. Sajda P. Machine learning for detection and diagnosis of disease // Annual Review of Biomedical Engineering, 2006, no. 8, pp. 537–565.
40. Suykens J.A., Vandewalle J. Least squares support vector machine classifiers // Neural Process Letters, 2004, vol. 9, no. 3, pp. 293–300.
41. Swets J.A. ROC analysis applied to the evaluation of medical imaging techniques // Investigative Radiology, 1979, no. 14, pp.109-21.
42. Tsay D., Patterson C. From machine learning to artificial intelligence applications in cardiac care: real-world examples in improving imaging and patient access // Circulation, 2018, vol.138, no. 22, pp. 2569–2575.
43. Twala B. An empirical comparison of techniques for handling incomplete data using decision trees // Applied Artificial Intelligence, 2009, no. 23, pp. 373–405.
44. Vapnik V., Lerner A. Pattern recognition using generalized portrait method // Automation and Remote Control, 1963, no. 24, pp. 774–780.
45. Ziyu Jin, Ning Li Diagnosis of each main coronary artery stenosis based on whale optimization algorithm and stacking model // Mathematical Biosciences and Engineering, 2022, vol.19, is. 5, pp.4568-4591. https://doi.org/10.3934/mbe.2022211
46. Zou K.H., O'Malley A.J., Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models // Circulation, 2007, is.5, vol. 115, pp. 654 – 657.

References

1. Ayvazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika: klassifikatsiya i snizhenie razmernosti* [Applied Statistics: Classification and Dimension Reduction]. M.: Finance and statistics, 1989. 607 p.
2. Belyalov F.I. *Kompleksnye problemy serdechno-sosudistykh zabolevaniy,* 2018, vol. 7, no. 1, pp. 84–93. https://doi.org/10.17802/2306-1278-2018-7-1-84-93
3. Vetrov D.P., Kropotov D.A. *Algoritmy vybora modeley i postroeniya kollektivnykh resheniy v zadachakh klassifikatsii, osnovannye na printsipe ustoychivosti* [Algorithms for choosing models and constructing collective solutions in classification problems based on the principle of stability]. Moscow, URSS, 2006, 112 p.
4. Vorontsov K.V. *Matematicheskie metody obucheniya po pretsedentam (teoriya obucheniya mashin)* [Mathematical methods of learning by precedents (the theory of machine learning)]. http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

5. Vorontsov K.V. *Lektsii po statisticheskim (bayesovskim) algoritmam klassifikatsii* [Lectures on statistical (Bayesian) classification algorithms]. http://www.ccas.ru/voron/download/Bayes.pdf
6. V'yugin V.V. *Matematicheskie osnovy teorii mashinnogo obucheniya i prognozirovaniya* [Mathematical foundations of the theory of machine learning and forecasting]. M., 2013, 387 p.
7. Duda R., Hart P. *Raspoznavanie obrazov i analiz stsen* [Pattern recognition and scene analysis]. M.: Mir, 1976, 511 p.
8. Zagoruyko N.G. *Prikladnye metody analiza dannykh i znaniy* [Applied methods of data and knowledge analysis]. Novosibirsk: IM SO RAN, 1999, 270 p.
9. Kardiovaskulyarnaya profilaktika 2017. Rossiyskie natsional'nye rekomendatsii [Cardiovascular prevention 2017. Russian national guidelines]. *Rossiyskiy kardiologicheskiy zhurnal*, 2018, no. 6, pp. 7-122. https://doi.org/10.15829/1560-4071-2018-6-7-122
10. Litvin A.A., Kalinin A.L., Trizna N.M. *Problemy zdorov'ya i ekologii,* 2008, vol. 18, no. 4, pp. 12-19.
11. Nevzorova V.A., Plekhova N.G., Priseko L.G. et al. *Rossiyskiy kardiologicheskiy zhurnal,* 2020, vol. 25, no. 3, pp. 10–16. https://doi.org/10.15829/1560-4071-2020-3-3751
12. Panev N.I., Evseeva N.A., Filimonov S.N., Korotenko O.Yu., Danilov I.P. *Meditsina truda i promyshlennaya ekologiya,* 2021, vol. 61, no. 6, pp. 365–370. https://doi.org/10.31089/1026-9428-2021-61-6-365-370
13. Samigulin T.R., Dzhurabaev A.E.U. *Nauchnyy rezul'tat. Informatsionnye tekhnologii,* 2021, vol. 6, no. 1, pp. 55–62. https://doi.org/10.18413/2518-1092-2021-6-1-0-7
14. Smirnova M.D., Svirida O.N., Fofanova T.V. et al. *Kardiovaskulyarnaya terapiya i profilaktika,* 2021, vol. 20, no. 6, p. 2799. https://doi.org/10.15829/1728-8800-2021-2799
15. Tu J., Gonzalez R. *Printsipy raspoznavaniya obrazov* [Principles of pattern recognition] / ed. Yu. I. Zhuravlev. M.: Mir, 1978, 411 p.
16. Chazova I.E., Oshepkova E.V. *Analiticheskiy vestnik,* 2015, no. 44(597), pp. 4-8.
17. Aravind Akella, Sudheer Akella Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. Future Science OA, 2021. https://doi.org/10.2144/fsoa-2020-0206
18. Breiman L. Random Forest. *Machine Learning*, 2001, vol. 45, no. 1, pp. 5-32.
19. Bordes A., Ertekin S., Weston J., Bottou L. Fast Kernel Classifiers with Online and Active Learning. *Journal of Machine Learning Research*, 2005, no. 6, pp. 1579–1619.

20. Cardiovascular disease. World Health Organization website. 2022. https://www.who.int/ru/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
21. Cortes C., Vapnik V. Support vector networks. *Machine Learning*, 1995, no. 20, pp. 273–297.
22. Cuocolo R., Perillo T., De Rosa E., Ugga L., Petretta M. Current applications of big data and machine learning in cardiology. *Journal Geriatric Cardiology*, 2019, vol. 16, no.8, pp.601-607.
23. Deo R.C. Machine learning in medicine. *Circulation*, 2015, vol. 132, no. 20, pp. 1920- 1930.
24. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, vol. 27, no 8, pp. 861-874.
25. Forgy E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 1965, vol. 21, pp. 768–769.
26. Foster K.R., Koprowski R., Skufca J.D. Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomedical Engineering Online*, 2014, no. 13, article no. 94. https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-94
27. Guyon, B. Boser, Vapnik V. Automatic capacity tuning of very large VC-dimension classifiers. In S. J. Hanson, J. D. Cowan, and C. Lee Giles, editors, Advances in Neural Information Processing Systems, Morgan Kaufmann, San Mateo, CA, 1993, pp. 147–155.
28. Ho T. K. The Random subspace method for construction decision tree forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, vol. 20, no. 8, pp. 832–844. https://doi.org/10.1109/34.709601
29. Karimollah Hajian-Tilaki Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, 2013, vol. 4, no. 2, pp. 627–635. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/
30. Khan Y, Qamar U, Yousaf N, Khan A. Machine learning techniques for heart disease datasets. *Proceedings of the 2019 11th International Conference on Machine Learning and Computing – ICMLC '19*, 2019. https://doi.org/10.1145/3318299.3318343
31. Khaled M.A. Prediction of heart disease and classifiers sensitivity analysis. *Almustafa BMC Bioinfirmatics*, 2020, no. 21, Article number: 278. https://doi.org/10.1186/s12859-020-03626-y
32. Kohavi R., Quinlan J.R. Decision tree discovery. C5.1.3. 1999. https://ai.stanford.edu/~ronnyk/treesHB.pdf
33. Kummar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics*, 2011, vol. 48, no. 7, pp. 277-89.

34. Lu Y., Dendukuri N., Schiller I., Joseph L. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Statistics in medicine*, 2010, no. 29, pp. 2532-2543.
35. Maheswari S., Pitchai R. Heart disease prediction system using decision tree and naive bayes algorithm. *Current Medical Imaging*, 2019, no. 15, pp. 712–77. https://doi.org/10.2174/1573405614666180322141259
36. Nefedov A. Support Vector Machines: A Simple Tutorial. 2016. https://svmtutorial.online/download.php?file=SVM_tutorial.pdf
37. Obermeyer Z., Ezekiel J.E. Predicting the future – big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 2016, vol. 375, no.13, pp. 1216–1219.
38. Quinlan J. R. Induction of decision trees. *Machine Learning*, 1986, vol. 1, no. 1, pp. 81-106.
39. Sajda P. Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering*, 2006, no. 8, pp. 537–565.
40. Suykens J.A., Vandewalle J. Least squares support vector machine classifiers. *Neural Process Letters*, 2004, vol. 9, no. 3, pp. 293–300.
41. Swets J.A. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 1979, no. 14, pp.109-21.
42. Tsay D., Patterson C. From machine learning to artificial intelligence applications in cardiac care: real-world examples in improving imaging and patient access. *Circulation*, 2018, vol.138, no. 22, pp. 2569–2575.
43. Twala B. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 2009, no. 23, pp. 373–405.
44. Vapnik V., Lerner A. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 1963, no. 24, pp. 774–780.
45. Ziyu Jin, Ning Li Diagnosis of each main coronary artery stenosis based on whale optimization algorithm and stacking model. *Mathematical Biosciences and Engineering*, 2022, vol.19, is. 5, pp.4568-4591. https://doi.org/10.3934/mbe.2022211
46. Zou K.H., O'Malley A.J., Mauri L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, 2007, is.5, vol. 115, pp. 654 – 657.

ДАННЫЕ ОБ АВТОРЕ

Павлова Анна Илларионовна, кандидат технических наук, доцент
Новосибирский государственный университет экономики и управления

ул. Каменская, 56, г. Новосибирск, 630039, Российская Федерация
annstab@mail.ru

DATA ABOUT THE AUTHOR

Anna I. Pavlova, PhD (technical sciences), Associate Professor
Novosibirsk State University of Economics and Management
56, Kamenskaya Str., Novosibirsk, 630039, Russian Federation
SPIN-code: 8714-1140
ORCID: https://orcid.org/0000-0001-6159-1439

Поступила 29.08.2022
После рецензирования 05.12.2022
Принята 23.12.2022

Received 29.08.2022
Revised 05.12.2022
Accepted 23.12.2022