IRSTI 28.23.25 DOI: https://doi.org/10.26577/JMMCS.2022.v115.i3.10

A.K. Berdaly , Z.M. Abdiahmetova*
Al-Farabi Kazakh National University, Kazakhstan, Almaty
*e-mail: zukhra.abdiakhmetova@gmail.com

PREDICTING HEART DISEASE USING MACHINE LEARNING ALGORITHMS

Increasing the accuracy of detecting heart disease is widely studied in the field of machine learning. Such study is intended to prevent large costs in the field of healthcare and is the reason for the misdiagnosis. As a result, various methods of analyzing disease factors were proposed, aimed at reducing differences in the practice of doctors and reducing medical costs and errors. In this study, 6 classification learning algorithms were used, including machine learning methods such as classification Tree, Close neighborhood method, Naive Bayes, Random forest tree, and Busting methods. These methods were collected by the University of Cleveland. Using heart.csv dataset, they were trained to make an effective and accurate prediction of heart disease. In order to increase the predictive capabilities of algorithms, all methods were trained primarily on non-standardized data. A study was conducted on how much data standardization affects the result using the Standard Scaler method. In the paper, this method helped algorithms such as KNN and SVC improve the result about 25%.

Key words: Classification, Standardization, Training Selection, Metrics, Busting, Confusion Matrix.

А.К. Бердалы, З.М. Абдиахметова*
Әл-Фараби атындағы Қазақ ұлттық университеті, Қазақстан, Алматы қ.
*e-mail: zukhra.abdiakhmetova@gmail.com

МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІ АРҚЫЛЫ ЖҮРЕК АУРУЛАРЫН БОЛЖАУ

Жүрек ауруларын анықтаудың дәлдігін арттыру машиналық оқыту саласында кеңінен зерттелуде. Мұндай зерттеу денсаулық сақтау саласында үлкен шығындардың алдын алу үшін және қате диагноздың қойылу себептерінен туындайды. Нәтижесінде дәрігерлердің тәжірибесиндегі айырмашылықтарды азайтуға және медициналық шығындар мен қателіктерді төмендетуге бағытталған ауру факторларын талдаудың әртүрлі әдістері ұсынылады. Бұл зерттеуде классификациялық оқытудың 6 алгоритмң, соның ішінде атап айтқанда жіктеу ағашы, жақын көршілер әдісі, аңғал Байес, кезейсоқ орман ағашы, бустинг әдістері қолданылды. Осы әдістерді Клевеленд университетінің жинақтаған heart.csv датасетіне қолдану арқылы жүрек аурулары бойынша машинаға тиімді және дәлдігі жоғары болатын болжам жасау үйретілді. Алгоритмдердің болжау қабілетін арттыру мақсатында барлық әдістер бірінші кезекте стандартталмаған деректерге оқытылды. Standart Scaler әдісін қолдану арқылы деректерді стандартизациялау нәтижеге қаншалықты әсер ететініне зерттеу жүргізілді. Зерттеу барысында бұл әдіс KNN мен SVC секілді алгоритмдерге нәтижені шамамен 25%-ға жақсартуға көмек беретіні анықталды.

Түйін сөздер: Классификация, стандартизация, оқыту таңдамалары, метрика, бустинг, шатасу матрицасы.

А.К. Бердалы, З.М. Абдиахметова*
Казахский национальный университет имени аль-Фараби, Казахстан, г. Алматы
*e-mail: zukhra.abdiakhmetova@gmail.com

ПРОГНОЗИРОВАНИЕ ЗАБОЛЕВАНИЙ СЕРДЦА С ПОМОЩЬЮ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Повышение точности выявления заболеваний сердца широко изучается в области машинного обучения. Такое исследование призвано предотвратить большие потери в здравоохранении и привести к неправильному диагнозу. В результате были предложены различные методы анализа факторов заболевания, направленные на снижение различий в опыте врачей и снижение медицинских расходов и ошибок. В данном исследовании были использованы 6 алгоритмов классификационного обучения, в том числе методы машинного обучения, такие как дерево классификации, метод ближайших соседей, наивный Байес, случайное лесное дерево, методы бустинга. Эти методы были обобщены университетом Клевеленда применяя к датасету ССЗ. Они были обучены делать эффективные и высокоточные прогнозы сердечных заболеваний. С целью повышения предсказательной способности алгоритмов все методы были обучены в первую очередь нестандартизированным данным. Проведено исследование того, насколько стандартизация данных с использованием метода Standard Scaler влияет на результат. В ходе исследования данный подход улучшил результаты алгоритмов как KNN и SVC почти на 25%.

Ключевые слова: Классификация, стандартизация, обучающая выборка, метрика, бустинг, матрица путаницы.

1 Introduction

Cardiovascular disease is a disease that poses a risk of death in the modern world and is the biggest problem, as predicted by medicine in terms of growth. According to World statistics, this disease is such a problem that it worries the whole world, which leads to a large mortality factor. According to the World Health Organization, about 20 million people die from heart disease. In England, cardiovascular diseases account for 34% of deaths, while in European countries these statistics reach 40%. According to the latest statistics, the number of deaths from cardiovascular diseases around the world is increasing, the main reason for this forecast is that the statistics of countries with the lowest risk of cardiovascular disease are increasing every year. But according to who forecasts, more than 75% of cardiovascular diseases can be prevented, thereby reducing the burden of developing diseases.

Purpose of the work: selection and description of machine learning methods in Big Data Processing, increasing accuracy in the process of big data learning and reducing machine learning time. Research objectives:

- analysis of the literature on the use of machine learning (ML) methods for data on heart failure;
- analysis of python language libraries and part of machine learning methods;
- initial analysis and pretreatment of data related to cardiac arrhythmias;
- use methods for classifying signs, selecting and filling in missing values;
- analyze obtained results;

- justification of the research results in the subject area.

Object of research: the object of research is the prediction of cardiovascular diseases using machine learning algorithms. Use methods that allow to study and analyze the data used to optimize the process of solving research problems. Using this method, to create a system based on predicting the disease, minimizing human participation in analysis and creating an optimal solution with the participation of machine learning algorithms.

2 Literature Review

Machine learning is an analysis method that allows us to conduct data training and analysis methods that we use to optimize the process of solving research problems. This method is a system based on minimizing human participation in analysis and creating an optimal solution with the participation of artificial intelligence intelligence systems. This article will explore machine learning methods to make predictions in the process of Big Data Processing and analyze some specific methods. Currently, there is an active implementation of machine learning methods in medical information systems (MIS). This is primarily due to the need to analyze a large amount of information about patients in real time, as well as predict whether to seek outpatient care or hospitalization within a given time frame [1]. For the database, there are many open sources for accessing acient records, and research can be conducted to use various computer technologies to identify this disease in order to make the correct diagnosis of the patient and prevent his death [2]. Patients are often diagnosed asymptomatic until death, and even if they are under supervision, trained personnel are required to detect cardiac abnormalities [3]. Heart disease was the cause of 6.2 million deaths between the ages of 30 and 70 in 2019 [4]. These diseases usually occur as a result of stroke, hypertensive heart disease, rheumatic heart disease, artery disease and other defects in the heart vessels and the heart itself [5]. In many countries, there is little experience in cardiovascular research and a significantly higher percentage of misdiagnosed cases, which can be solved by developing accurate and effective methods for predicting heart disease at an early stage through analytical support for clinical decision-making through digital medical records [6].

Amin Ul Hak, Jiang Ping Li, Muhammad Hammad Memon, Shah Nazir and Ruinan Sun were tested on their systems in a Cleveland heart disease dataset. Seven well-known classifiers, such as logistic regression, KN, AN, SM, NB, DT and random forest were used with three algorithms for selecting functions Relief, mRMR, and LASSO, which are used to select important functions. In terms of features SVM (linear) with the selection of functions, the performance of the mrmr algorithm was better than that of other classifiers [7]. Fajr Ibrahim Alarsan and Mamun Yunets received a data set of 205.146 lines, which were randomly divided into two parts: training and testing. They compared the Random Forest and Decision Tree Classifier algorithms in machine learning of this data set. In a random forest, the learning process is faster than in a decision tree and in a decision tree, the testing process is faster than in a random forest. The parameters of both algorithms were changed manually. The optimal values for the configured parameters could be obtained by running cross-checking methods, but the algorithms took a lot of time [8]. Jiang Yi, Zhang X, Ma R, Wang X, Liu J. , Kerman M, Yang Yi, Ma J, Son Yi, Zhang J. , He J, Go C, Go X chose dataset as the data that monitored 1,508 Kazakh subjects in China at the initial level without cardiovascular diseases. All subjects were randomly divided into a study sample (80%) and a test sample (20%). LR, SVM, DT, RF, KNN, NB and XGB were used to predict outcomes in cardiovascular diseases. LR and SVM had better predictive characteristics than other machine learning models in the context of discrimination and calibration. LR was similar to the predicted effectiveness of SVM in predicting the risk of cardiovascular diseases and surpassed other ML models. The sensitivity of LR was higher than that of SVM and the specificity gave the opposite result [9].

3 Problem Settings

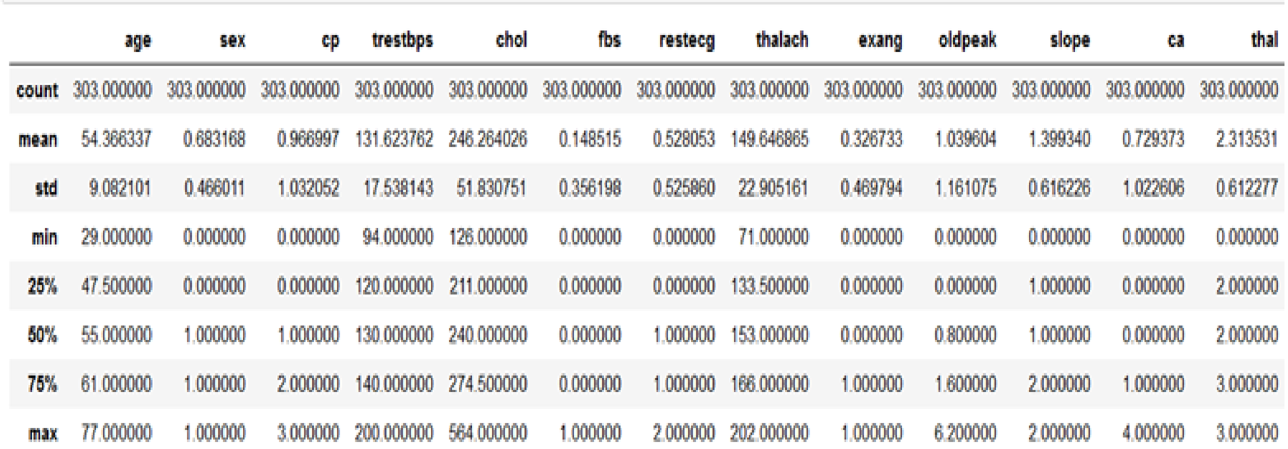
This section discusses sorting data from the collected databases, conducting pre-machine learning processing measures, fully studying target variables, dividing them into machine learning and testing stages and learning these information using machine learning method classifiers. Through the selected classifiers, the level of training is evaluated and measures are taken to improve the results. The first step is to access the database used in data training. The dataset taken from the database consists of 14 columns of 303 consecutive factors affecting the symptoms of cardiovascular disease. This database was collected by the Cleveland Clinic, which was connected with the university clinics of Zurich and Basel. The database originally consisted of 72 columns, and as a result of removing columns that did not attach much importance to special processing and research activities, 14 columns were left.

Table 1. Database analysis

Column name	Meaning	Range
Age	Patient age	[29, 77]
Sex	Gender	0 = female, 1 = male
Cp	Type of pain	[1, 4]
Trestbps	Blood pressure in a calm state	[94, 100]
Chol	Cholesterol levels	[126,564]
Fbs	Glucose levels in the blood of a hungry person	0 = false, 1 = true
Restecg	ECG(Electrocardiography)	[0, 2]
Thalach	Maximum pulse rate	[71, 202]
Exang	Angina pectoris during physical exertion	0 = no, 1 = yes
Oldpeak	St depression level	[0, 6.2]
Slope	ECG at maximum load	[1, 3]
Ca	Fluorescent color important blood vessel number	[0, 3]
Thal	A type of blood disease called thalassemia	3, 6, 7
target	Heart disease	0 = no, 1 = yes

We can show statistical characteristics for numeric attributes in the database. Statistical values are represented as the total number of attributes, the average, standard statistical deviations, the smallest and largest values, as well as indicators of 25%, 30% and 75% on 3 quartils. You can see it in the table below.

Table 2. Statistical indicators for databases



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000

Since the indicator of people with cardiovascular diseases was taken as a target variable, the indicators for this variable were visually displayed. Age indicator of the number of people suffering from heart disease according to Figure 1. As we have seen, the most sick people can be called the age range of 40 to 55 years.

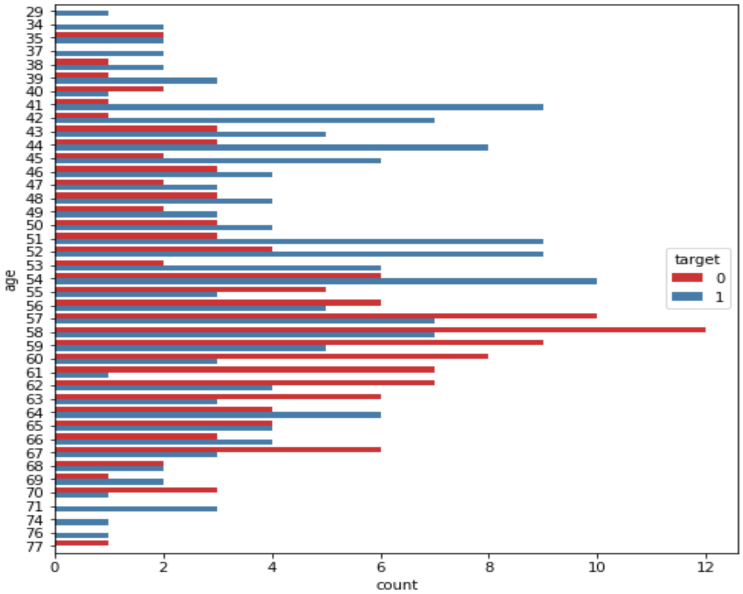


Figure 1: Quantitative indicator for the target variable

4 Materials and Methods

To check the accumulated commands, we first look at whether there are zero elements in the dataset, and if such data is found, fill in the spaces by calculating the median or average value of this column. Disable them because dictionary columns are not involved in training. Algorithms of the machine learning method are used by setting target variables.

The general picture of the work carried out on the methodology is shown in Figure 2:

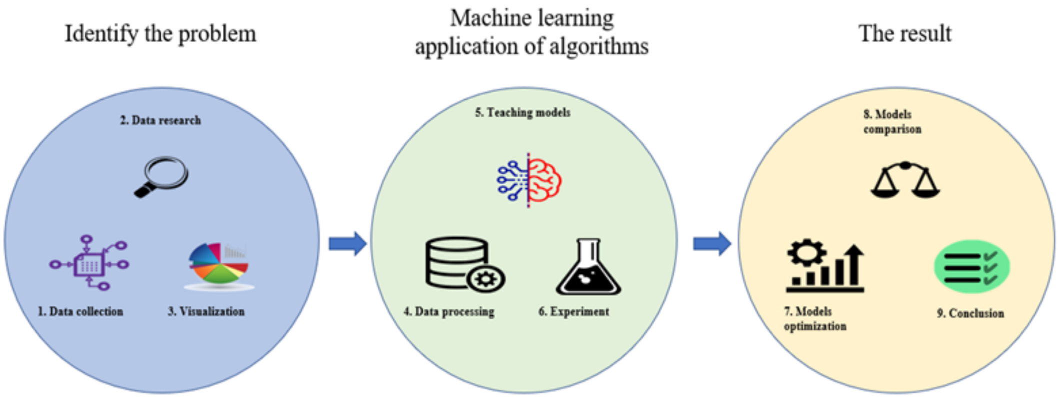


Figure 2: General research plan

Until measures to improve the accuracy of the algorithms used give good results, it is necessary to implement such measures as training, avoiding mistakes in the course of excessive training. Classification techniques used to detect cardiovascular diseases are as follows: Decision Tree Classifier, Kneighbors Classifier, Logistic Regression, XGBClassifier, Random Forest Classifier, Support Vector Classfier. Although training is carried out using such techniques, cross – checking with the target variable column of the dataset is carried out in order to increase the result. Cross-validation (verification) is the process of improving the result of an algorithm by training each time with different random values with a random transfer of a target variable to a test set in order to improve the learning efficiency of machine learning operations.

The classifier tree method is a tree-type structure consisting of certain rules that represent the result in the learning process in a hierarchical type. It consists of 2 types of elements, a node and a leaf. Elements to be written in the node if the elements that affect the value of the target variables are written in the Leaf, the functions of the target variables are written in the Leaf. The decision tree is often used because it gives good results in statistical reports, including in medical reports for more probabilistic reports, by classifying the same data, making better predictions, clarity, and simple processing of the data without converting or causing severe distortion [10].

The k nearest neighbors method is an algorithm for classifying objects by class by dividing them into groups previously distributed by region after calculating the distances by weight by vote. This method is considered the simplest of the classification algorithms. It is a classifier algorithm that can be used in cases where there is little information about the data in the preliminary data separation, and is completely unknown. The optimal method for understanding and implementing KNN. Therefore, this situation should be taken into account for any calcification calculations. Its main advantages are that the process is very clear, time-consuming, efficiency and accuracy are often high, and there are methods that eliminate noise in the process that work only for KNN [11].

The logistic regression method is a statistical method that classifies a classification using a linear classification line. The main idea is to determine the optimal line that best divides data through a set of data. The range of logistic regression covers the range from 0 to 1. In addition, this method does not require a connection between input and output data. Logistic regression is a method in medical research that allows you to perform several tests at the same time, minimizing external factors. If the model structure created by the researcher avoids raw data,then the probability of logistic regression is also high [12].

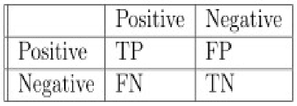
The XGBoost method is a method that belongs to the ensemble method, designed to improve gradient descent, with optimal and high accuracy. This is a method that aims to get the best results by training multiple decision trees in parallel to improve the gradient. Through XGBoost, trees grow rapidly and parallel trees are erected, the final decision is made by an ensemble voice. In this method, random forest trees and decision trees are solved by using models and making comparisons with their parameters [13].

The random forest tree method is another type of algorithm that uses the ensemble method. An algorithm that randomly creates a forest of decision trees, takes forest trees of different selections, matches them to the classifier, and finally takes the average value in order to increase accuracy. The main advantage is the ability to achieve good results when working with large groups and classes, independence from the scale of learning, and the ability to perform high parallelization. Therefore, the random forest tree is an effective predictor [14].

The method of reference vectors is a set consisting of intensive learning algorithms and bringing changes through hyperactivity to a single norm. The idea of the method is that we place data elements consisting of points on the n-dimensional plane, creating hyperspace by creating a classification that best defines classes. SVM also has a core, which is used to convert data entered into the plane by the cores to a large one, taking it as small. It is mainly used for Tex cataloging, recognizing handwritten numbers, finding tones, classifying images, and gene expression using a microchip [15].

In the future, we will use assessment metrics to evaluate the training of these 6 used methods. To do this, in the process of comparing the algorithm-trained results of y with the true value of y in the target variable, a reflection matrix is created, as in Table 3. The result of algorithms based on the generated matrix will be evaluated.

Table 3. Confusion matrix



	Positive	Negative
Positive	TP	FP
Negative	FN	TN

True Positive (TP) – the classifier assumes that the positive result is positive.

True Negative (TN) - the classifier assumes that the negative result is negative.

False Positive (FP) - the classifier incorrectly predicts a negative result as positive.

False Negative (FN) - the classifier incorrectly predicts a positive result as negative.

The classifier is evaluated using the formulas of the metrics listed below:

Accuracy – the total accuracy of the model, the amount of accuracy of classifiers when compared with the main values.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is an indicator that the classifier finds positive and is actually positive.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is an indicator of true positive classes among all positive classes found by the algorithm.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-score is the hormonal average of accuracy and completeness.

$$F1 - score = \frac{2TP}{2TP + FP + FN} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Results

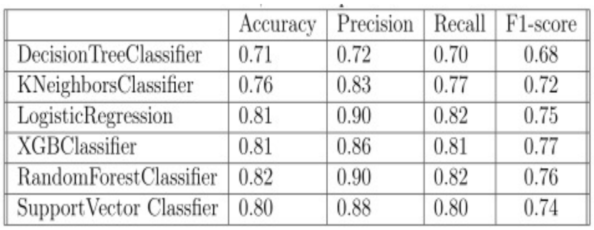
Data source of a 303-row, 14-column collected by Cleveland Medical Center for cardiovascular disease has gone through processing measures that consist of many steps. As a result of the processing measures, no particularly strong outs and zero elements were found in the database. The absence of columns that strongly influence each other on the data was observed through the correlation matrix. After processing, 30 percent of the data was sent for training. Subsequently, 10 classification algorithms were trained. It includes algorithms Decision Tree Classifier, Kneighbors Classifier, Logistic Regression, XGBClassifier, Random Forest Classifier, Support Vector Classfier.

During the training of each algorithm, the result was increased by standardization using the Standard Scaler function. The Standard Scaler function tries to show good results by normalizing our data so that the average value does not exceed 0 and the standard deviation does not exceed 1, which gives the opposite effect before applying the algorithm. Algorithms such as Decision Tree Classifier showed a decrease in accuracy from 0.7142% to 0.7023% from the scattered neural structure algorithm, while Kneighbors Classifier helped to increase the accuracy from 0.5934% to 0.7692%. The same result was obtained from the support Vector Classfier algorithm, which increased the accuracy from 0.5604% to 0.8021% by standardization.

In addition to standardization, we tried to find the most optimal parameters and increase the result using the GreedSearchCV algorithm. GreedSearchCV refers to a cross – validation operation. It is one of the most powerful tools in machine learning, the main reason for which the correct choice of parameters is the main guarantee of good results. If the parameters are chosen correctly, then, of course, the training will also go well. As for work, it calculates the result for each parameter over the entire connection, providing us with the best indicator. The result was not satisfied, there were significant delays in terms of time, and the result of the algorithm did not show much difference from standardization.

Thus, 6 algorithms were evaluated on 4 metrics. The results were compared among themselves. This can be seen in Table 4. The best indicator for the Accuracy metric was the result of the Random Forest Classifier algorithm.In Figure 3, dynamic comparisons were made. You can clearly see the real difference through the diagram.

Table 4. Comparison of results of classifiers



	Accuracy	Precision	Recall	F1-score
DecisionTreeClassifier	0.71	0.72	0.70	0.68
KNeighborsClassifier	0.76	0.83	0.77	0.72
LogisticRegression	0.81	0.90	0.82	0.75
XGBClassifier	0.81	0.86	0.81	0.77
RandomForestClassifier	0.82	0.90	0.82	0.76
SupportVector Classfier	0.80	0.88	0.80	0.74

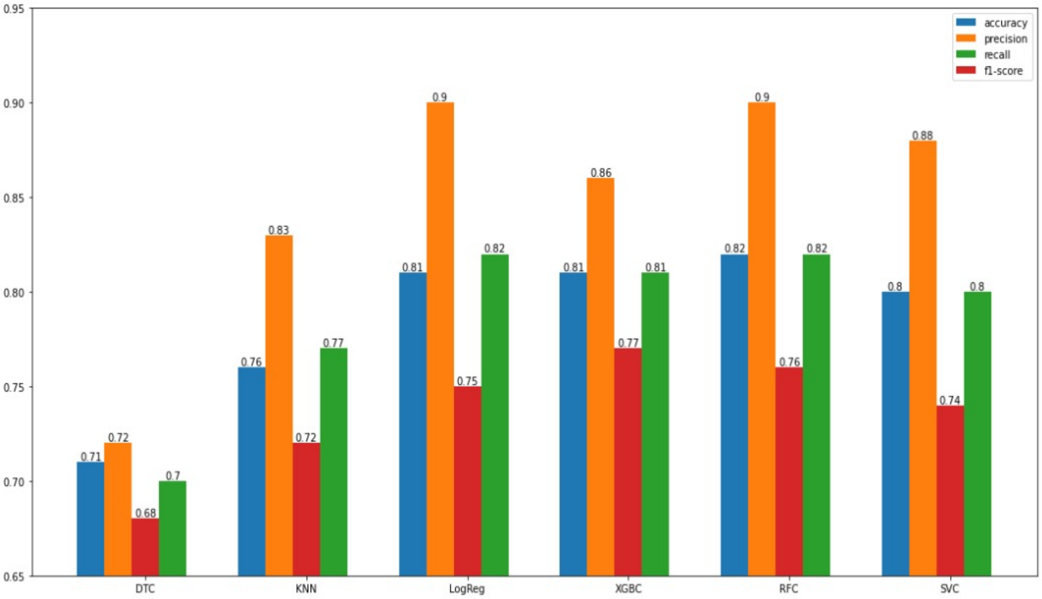


Figure 3: Comparison of results of classifiers

Conclusion

Heart disease is one of the main problems of society, as the number of people with heart diseases is increasing day by day. The growth of Statistics is influenced by many factors, such as the time spent by medicine to predict diseases or the lack of an accurate diagnosis. It is difficult to manually determine the probability of heart disease based on many such factors. But with deep data analysis and machine learning models, it is possible to identify diseases and treat these diseases in a timely manner. For this purpose, relevant data on heart disease collected by the University of Cleveland were studied. Work achieved and done during the study:

- analysis of the literature on the use of machine learning (ML) methods for data on heartbeats was carried out;
- analysis of python language libraries and part of machine learning methods;
- primary analysis of data on heart beauties and pre-processing;
- the marks were stitched, selected and methods of filling in the missing values were used;
- the results obtained were analyzed;
- based on the results, a comparison was made between the models.

According to the conducted research, the classification method showed the highest results. Its metrics showed accuracy = 0.82%, precision = 0.91%, recall = 0.83%, and f1-score = 0.76%. In the future, training of the algorithm on various data will continue, increasing these results given by Random Forest. Further experiments are developed on algorithms and optimal solutions are developed using various methods. Algorithms that have been trained to read various data on heart disease are also good at making predictions.

References

[1] Gusev A .V., Novitskiy R.E., Ivshin.A.A."Machine learning based on laboratory data for disease prediction", *FARMAKOEKONOMIKA. Modern Pharmacoeconomics and Pharmacoepidemiology*(2021): 8-9 p.

[2] R. Bharti [et al.],"Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Computational Intelligence and Neuroscience*(2021): 3-6 p.

[3] Dimitris Bertsimas, Luca Mingardi, Bartolomeo Stellato "Machine Learning for Real-Time Heart Disease Prediction", *IEEE Journal of Biomedical and Health Informatics* (2021): 12-14 p.

[4] Mensah G.A., Roth G.A., Valentin Fuster, "The Global Burden of Cardiovascular Diseases and Risk Factors: 2020 and Beyond ", *Journal of the American College of Cardiology,* (2019): 2-9 p.

[5] Amini M., Zayeri F., Salehi M., "Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017", *BMC Public Health,* (2021): 12 p.

[6] Md Mamun Ali [et al.], "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison", *Computers in Biology and Medicine.* (2021): 6-7 p.

[7] Amin Ul Haq [et al.], "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms", *Mobile Information System* (2018): 11-13 p.

[8] Fajr Ibrahem Alarsan, Mamoon Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms" , *Journal of Big Data* (2019): 8 p.

[9] Yunxing Jiang [et al.], "Cardiovascular disease prediction by machine learning algorithms based on cytokines in kazakhs of china" , *Clinical Epidemiology* (2021): 5 p.

[10] Yan Yan Song, Ying Lu., "Decision tree methods: applications for classification and prediction" , *Shanghai Archives of Psychiatry* (2015): 1-7 p.

[11] Pбdraig Cunningham, "K-Nearest Neighbour Classifiers-A Tutorial" , *ACM Computing Surveys* (2021): 6-9 p.

[12] Sandro Sperandei, "Understanding logistic regression analysis" , *Biochemia Medica* (2014): 5-9 p.

[13] Candice Bentejac, Anna Csurgo, Gonzalo Martinez-Munoz, "Data classification using support vector machine" , *Artificial Intelligence Review* (2021): 2-7 p.

[14] Leo Breima, "Random forests" , *Machine Learning 45,5-32* (2001)

[15] Durgesh K Srivastava, Lekha Bhambhu, "Data classification using support vector machine" , *Journal of Theoretical and Applied Information Technology* (2010): 3-9 p.