

Development of Heart Disease Diagnosis Concept Using Machine Learning

Bohdan Koval^a, Iulia Khlevna^a and Svitlana Shabatska^b

^a Taras Shevchenko National University of Kyiv, Volodymyrska str., 60, Kyiv, 01033, Ukraine

^b Bogomolets National Medical University, T. Shevchenko boulevard, 13, Kyiv, 01601, Ukraine

Abstract

The work is devoted to the development of a classification model using machine learning to be used in the applied industry of healthcare. It establishes prospects of modern information processing methods for medicine. The research was conducted on the basis of patients' medical records statistics to diagnose heart diseases. It proposes a conceptual approach to the application of machine learning to solve this problem, and forms the tools and the approach to a determination of connection between medical data records attributes and formation of results (diagnostics). The existing expert systems for the detection and diagnosis of heart diseases have been processed and analyzed. The work includes a research analysis of a set of medical data, which may be taken as a basis of operations and calculations with data that requires transparency and predictability. It constructs examples of machine learning models for heart disease detection with classification accuracy in the range of 80-89%. It gives the analytics and explanations of the received models, calculates a diagnosis outcome and what factors and to what extent influence it, therefore the models are not black boxes and can be clearly examined by clients (patients and doctors). The work outlines the perspectives and potential ways of development of the formed data preparation, analysis, and research concept.

Keywords ¹

machine learning, classification, anomalies detection, data visualization, data mining

1. Introduction

A large number of modern practical problems belong to the class of constraint satisfaction problems [1]. The target functions in such tasks are, as a rule, undifferentiated and (or) poly-extremal dependencies. The use of classical methods of continuous optimization and in many cases discrete optimization is impossible [2]. Combinatorial optimization methods, evolutionary algorithms, etc - are used to solve such tasks. The functional dependencies can be set tabularly or algorithmically. In this case evolutionary algorithms are most often preferred.

Exactly, the use of this algorithms does not require strict target functions constraints, but does not guarantee the finding of a global optimum, although according to certain conditions there is a probability convergence. The obtained solutions are considered suboptimal.

Historically, the first methods of evolutionary optimization were genetic algorithms and evolutionary strategies [3, 4]. These methods allowed to consider optimization problems differently and expanded the subject base of optimization technologies. Also, these methods were based on the ideas of natural evolution. In particular, genetic algorithms traditionally use the principle that the best parents tend to have better children.

Two parental potential solutions are involved in generating potential solutions-offspring. In evolutionary strategies, potential offspring solutions are generated around a single parent solution. This

Information Technology and Implementation (IT&I-2021), December 01-03, 2021, Kyiv, Ukraine

E-MAIL: bohkoval@gmail.com (A.1); yuliya.khlevna@gmail.com (A.2); sveta.shabatska@gmail.com (A.3)

ORCID: 0000-0002-3757-0221 (A.1); 0000-0002-1807-8450 (A.2); 0000-0002-9885-5546 (A.3)



© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

is the main difference between genetic algorithms and evolutionary strategies in the generation of offspring solutions.

Our hypothesis is that more potential parent solutions that will be involved in generating potential successive solutions will improve the accuracy of the solutions and speed up the convergence of the optimum search algorithms. This progress will be achieved by in-depth study of potential promising solutions and by reducing the number of algorithm steps in non-perspective directions.

2. Theoretical base and features of heart diseases detection and analysis using machine learning

2.1. Definition of the subject area of research

Of all the applications of machine learning, diagnosing any serious disease using complex and complex mathematical models (which are mostly black boxes, because it is difficult to justify what are the real factors behind a particular parameter and coefficient) is one of the most difficult tasks. [5]. If the result of a model is a specific course of treatment (possibly with side effects), surgery, or no treatment, people want to know why that particular result was obtained.

Electrocardiography, echocardiography, myocardial images using nuclear medicine, bicycle ergometry and carotid artery ultrasound are used to diagnose cardiovascular diseases. One of the biggest difficulties in working with medical indicators is the data: their volume, structure and presentation [6].

Since the early eighties of the last century, the so-called Medical Information Systems (MIS) - systems of workflow automation in treatment and prevention facilities, which combine a decision support system for medicine, electronic medical records, digital research data, financial and administrative information - began to spread rapidly. We can only imagine how fast the annual data growth in such systems is. In addition to the extremely large volume, the records in the system have other features that prevent analysis: gaps in the data, inaccuracies or subjectivity of judgment. Thus, there is a new term that we associate with the data subject to these characteristics - Big Data [7].

Obviously, without innovative high-tech data analysis tools, the probability of error is very high, because it is easy to miss an important factor, because attention was paid to another, less critical. Specially trained algorithms are able to find connections and dependencies that previously went unnoticed. An important aspect that distinguishes medical data from most others is that the quality, objectivity, timeliness and accuracy of the results are critical and the result must be constantly substantiated. Difficulties are also caused not only by the data itself, but also by the process of collecting them. All data comes from different providers, so the data is often cluttered and unstructured. Dirty data can quickly nullify complex analytics, especially when looking at information from different resources that can record data in a variety of formats.

2.2. Analysis of methods for heart disease detection and diagnosis

The healthcare information technology industry is changing daily. Therefore, organizations that have not entered the market for some time may not be aware of the latest innovative proposals from ambitious new "players" in the industry. As a result of rapid development, a new culture of scientific advances in machine learning is emerging, enabling the healthcare sector to take advantage of a number of revolutionary tools that use natural language processing, pattern recognition and in-depth training to provide more qualified services. The market for data mining technologies based on the solutions they try to offer can be divided into 4 main categories.

1. Analysis of pathology images. Improving the analysis of pathology images through machine learning is of particular interest to healthcare organizations. Machine learning can complement the skills of radiologists by showing more serious changes in faster scanning of images, which can potentially lead to earlier and more accurate diagnosis of diseases. A number of technology manufacturers have already begun to invest heavily in image analysis and analysis projects.

IBM Watson is deploying a clinical imaging service to help identify aortic stenosis, while Microsoft has focused on biomarker phenotyping research to complement its cancer research [8].

Scientific institutions are also part of the first level of modern pattern recognition. At Indiana University and Purdue University, Indiana, researchers are uncovering machine learning algorithms on pathology images to predict the recurrence rate of acute myelogenous leukemia. [9].

At Stanford University, machine learning tools have shown better results than experts, distinguishing between two types of lung cancer. The computer has also outperformed its competitors in predicting patient life expectancy [10].

2. Processing of natural language and text data. Unstructured data is everywhere in the healthcare field, whether it was created on purpose or accidentally.

Lab reports images in PDF format, voice recordings of consumer interactions, and text input with e-mail are important issues for traditional analysis tools. Machine learning offers a new way to derive value from these data sources. Using natural language processing (NLP) algorithms, machine learning algorithms can convert text images into documents, edit them, extract semantic values from these documents, or process user searches in plain text to increase the accuracy of the results.[11].

3. Decision support and forecasting. The ability to extract values from large amounts of free text is also crucial to support clinical decisions and prediction - another area where the development of machine learning illuminates the road. Prompt detection and elimination of risks can significantly improve outcomes for patients with any number of serious diseases, both clinical and behavioral.

4. Disease diagnosis using machine learning methods. The medical field has always been and remains one of the most important areas, into the development of which mankind puts a lot of effort. This is primarily due to the high cost of error, because we are talking about human life. 90% of the success of disease treatment depends on timely and correct diagnosis. For a long time, this responsibility fell on the shoulders of doctors. But even the most experienced professionals are, above all, people. First, a person cannot know everything, and second, there is a possibility of cognitive error due to fatigue, stress or inattention. A separate number of problems are caused by medical data: their volume, quality and presentation format. More than 80% of cognitive errors are caused simply by misinterpretation of data. Here come to the aid of the latest technologies that use artificial intelligence and machine learning as a diagnostic tool.

The basis of medical diagnosis is the task of classification. The diagnosis can be reduced to a problem with displaying data in one of N different results. In some cases $N = 2$: the task is only to determine whether the patient has a certain disease (1) or not (0) on the basis of data (e.g. X-ray or cardiogram). It is worth noting that, as a rule, preference is given to algorithms that the output has not only the answer to which class the object belongs, but also the probability that the object belongs to each class [12, 13]. From the conducted analysis it is established that it is appropriate to form the concept of diagnostics of heart diseases, developing methods of machine learning. In particular, algorithms based on the knowledge and analytical abilities of one or more experts in the field of medicine with the formation of a system that is able to draw logical conclusions based on this knowledge, thereby providing conclusions in the diagnosis of heart disease.

2.3. Indicators of a heart disease detection and analysis in the expert system

Medical information system (expert system in medicine) in the narrow sense - a set of technical means and mathematical software designed to collect, analyze medical and biological information and issue results in a user-friendly form. Depending on the type of tasks to be solved, MIS can be divided into three groups: information and reference, information and logic, and management. The development of any expert system in medicine should be based on some principles of building systems, the study and formulation of which is engaged in systems engineering. The use of knowledge of modern medicine is of paramount importance in the development of MIS. When developing a specific diagnostic or prognostic system, it is necessary to consistently address a number of issues:

- 1) choice of purpose and determination of the main purpose of the system;
- 2) selection of the system structural scheme;

3) compiling a list of nosological forms and collecting statistically reliable information on the symptoms of conditions;

4) construction of a decisive rule for solving problems of evaluation of medical information and issuance of conclusions on diagnosis and forecasting, i.e. development of the algorithmic basis of the system.

Data and knowledge processing is reduced to three main stages. At the first stage, the elements of information are placed in certain structures - databases and knowledge bases. At the second stage, the database and the knowledge base are streamlined: their structure changes, the order of placement of information, the nature of the relationships between the elements of information. At the third stage carry out operation of a DB: search for the necessary information, decision-making, editing of databases and knowledge [14]. There are different algorithms for medical diagnosis. The simplest algorithm, which often does not require the use of computers, is the so-called "tabular algorithm", based on the use of special diagnostic tables. Other, potentially more effective, types of medical diagnostic algorithms:

1) probabilistic method, which consists in calculating the probabilities of diseases according to the Bayesian formula;

2) Wald's method of sequential statistical analysis;

3) method of searching for clinical precedent;

4) phase interval method;

5) logical basis method;

6) algorithms based on neural networks.

A fundamental feature of any, even with the use of powerful technical means, diagnosis: to achieve 100% reliability of the diagnosis is almost impossible due to the fundamentally difficult reasons of both objective and subjective nature. Causes of an objective nature - individual characteristics of the organism, atypical course of the disease, errors in measuring the signs, overlapping areas of the space of signs. Causes of a subjective nature are the doctor's mistakes due to a number of factors [15].

Expert systems are not only a technology, but also a philosophy of using medical knowledge. The most useful feature of the expert system is that it uses high-quality experience of medical professionals to solve the problem.

2. Development of heart disease detection and analysis models using machine learning

2.1. Indicators formation for detection and analysis of heart disease in the expert system

A set of data on medical indicators of patients diagnosed with heart disease will be used for analysis. The data set [16] consists of appeals to a medical institution that has agreed to participate in the collection of relevant data. Therefore, the set includes data from both patients with confirmed heart problems and those diagnosed with healthy ones. The database included basic indicators of the patient (age, sex, etc.) and specialized medical (cardiogram results, etc.). Each record also contains an indicator (flag), which indicates whether the patient's final heart problems are diagnosed or not, and the severity of the disease on a scale from 0 to 4, where 0 - no disease, 4 - severe heart disease.

The use of such data sets with machine learning algorithms can help to create models for diagnosing and analyzing the occurrence and course of heart disease in patients, which in the future may contribute not only to more effective treatment but also prevention and prevention of these diseases, which can be expected to improve quality. and life expectancy. Also, around the created models it is possible to build a service - an information system, which can be joined by medical institutions, individuals, etc.

The data set contains a number of variables, the target of which is the state of the presence or absence of heart disease. Names and numbers of patients' social insurance were removed from the database and replaced with fictitious values. Dataset is a raw set of binary data, that will be used for the heart disease diagnosis of patients. The initial data set (rawBinaryData) contains 77 values (property) for each record (entity). The essence of the entity is the diagnosis of the patient, i.e. for each treatment and subsequent

diagnosis of the patient a separate record is created, which includes patient data and diagnostic results. Thus, each record is characterized by the patient's name and 76 attributes.

In the final data set (tidy dataset), which will be used for further work, it is necessary to leave only the attributes that may be relevant to our model and normalize / prepare them.

In the final data set we will expect attributes in the normalized form (number # indicates the ordinal number of the attribute in the initial data set):

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (attribute to diagnose)

The algorithm of initial data processing is:

1. We transform binary data into a text representation by decoding the binary representation using 'unicode_escape' encoding to avoid conflicts.

2. Replace \n with “ ” to place the data on the same order of attachments (so that there are no gaps in the records).

3. At this stage, the records are placed in one line and separated by a name. Therefore, to place each entry on its own line, replace 'name' with \n. (name is a constant that does not tell us anything).

4. We create dataframe from raw data: we initialize the constructor with the received csv-like text representation of the data, as a result we receive the normalized dataset of initial data.

5. Change the values of categorical quantities to improve their interpretation.

6. Check the data types. Some of them are not true (for example, the categorical values that we changed in the previous step), we adapt them.

7. We need to create dummy variables for categorical variables. For example, instead of having "men" and "women", we will have a "man" with values of 0 or 1 (1 is a man and 0 is therefore a woman). Normalizing thus a dataset, we receive the final dataset - tidyDataset.

As a result, we got tidyDataset, saved in the file tidy_dataset.csv. The data set is ready for further work, creation of models based on it and use in machine learning.

2.2. Modeling of heart diseases detection and analysis using machine learning

In total, the system implements 6 stages of diagnosing heart disease in a patient. The information system will help diagnose, analyze and predict users' heart disease. Below are presented and described all 6 stages of system operation:

1. Signal receivements. The system receives a request to process patient data through the ARI, which can be connected to smart watches, medical systems, mobile devices. The connection is encrypted to protect the client's private data.

2. Data Processing. The system examines the obtained data, normalizes them and uses data mining algorithms to extract useful data from the obtained ones. This helps the system to operate on what is really important and ignore the so-called white noise.

3. Building of mathematical models. Based on the prepared data, using machine learning algorithms, we create and train models for the detection and prevention of heart disease.

4. Results formation. The system tests the obtained models, selects the most accurate and based on them forms the final result of the diagnosis of the patient's heart disease. A comprehensive diagnostic analytics is being created.

5. Notification of results to a client. The client uses the web interface of the mobile device to receive diagnostic results. The patient can also indicate other people who should get the result, e.g. your doctor.

6. Providing a report and recommendations for further action. On the screen of the device the client is shown the complex report on the received diagnostics and recommendations are given.

To implement the software product, the Python programming language and the ecosystem around it were chosen. In combination with Python we will use the libraries implemented on its basis:

- pandas - data processing and analysis, the library provides data structures and operations for table and time series manipulation, it is optimized for working with large amounts of data;
- numpy - adds support for large, multidimensional arrays and matrices, along with a vast collection of high-level mathematical functions for operations with them;
- matplotlib - offers an object-oriented software interface for inserting various visualizations;
- scikit-learn - introduces various algorithms for solving problems of classification, regression analysis, clustering, including the method of support vectors, K-nearest neighbors, random forests and others;
- xgboost - framework to work with gradient boosting algorithms [13].

We start with the research of the data set and identify patterns in it. Taking into account empirical data on risk groups, we investigate the distribution by age and heart rate. As we can see on Fig. 1, there are several age categories that are at risk. Interestingly, the statistics are not increasing with age, and the most problematic ages are 41-45 years and 51-54 years.

From Fig. 2 we can see that most people of this age also have an increased heart rate compared to peers. During the study of the data it was found that it is effective to create fictitious variables based on categorical, and therefore convert the variables `cp`, `thal`, `slope`:

```
dummyCp = pd.get_dummies(df['cp'], prefix = "cp")
dummyThal = pd.get_dummies(df['thal'], prefix = "thal")
dummySlope = pd.get_dummies(df['slope'], prefix = "slope")
```

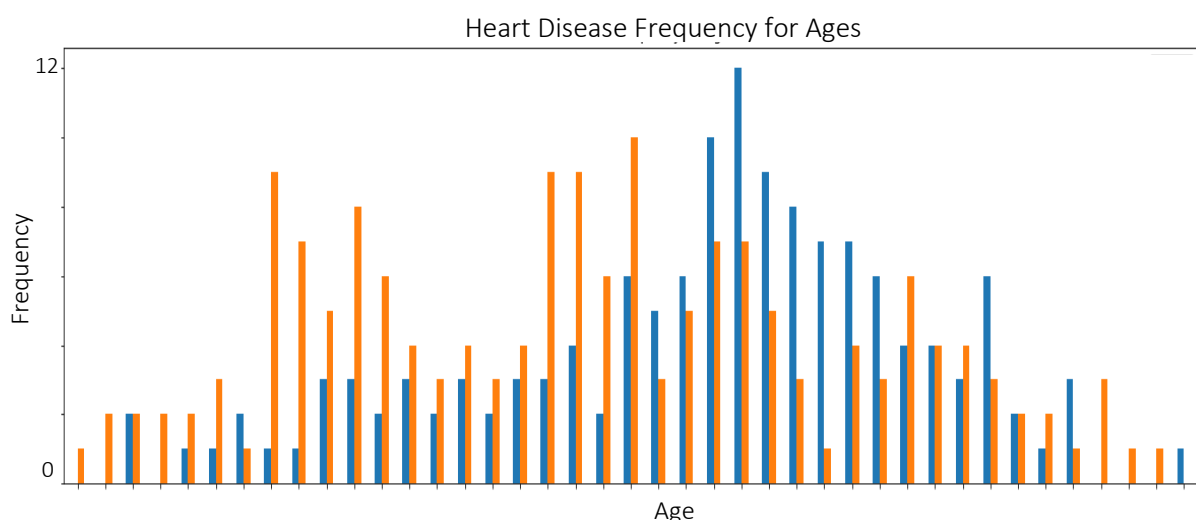


Figure 1: Distribution of heart diseases diagnosis by age (0 - no problems detected, 1 - problems detected)

Let's start with logistic regression. We can create and initialize the model ourselves, but for convenience we use the scikit-learn package and import logistic regression from it:

```
accuracies = {}
```

```
lr = LogisticRegression()
lr.fit(x_train.T,y_train.T)
acc = lr.score(x_test.T,y_test.T)*100
```

We will demonstrate the learning process of the model and the obtained result (Fig. 3).

We derive the result of the model - 84.39% accuracy. We initialize and train the model of K-nearest neighbors:

```
scoreList = []
for i in range(1,20):
    knn2 = KNeighborsClassifier(n_neighbors = i
    knn2.fit(x_train.T, y_train.T)
    scoreList.append(knn2.score(x_test.T, y_test.T))
```

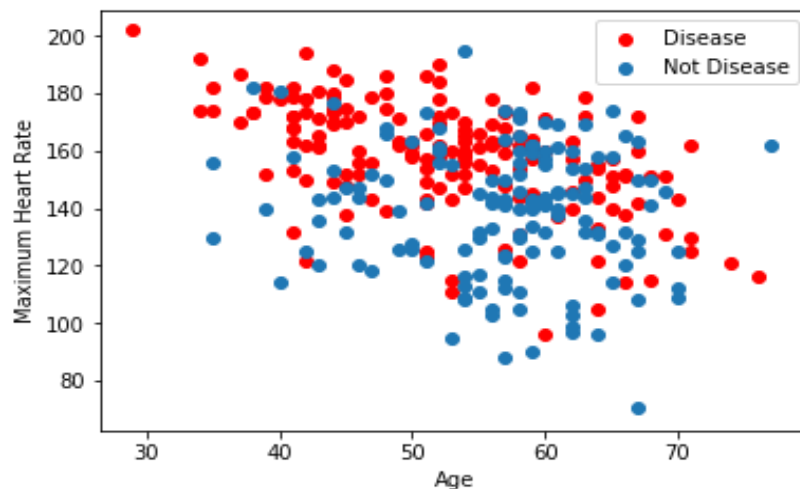


Figure 2: Distribution of heart disease diagnosis by maximum heart rate

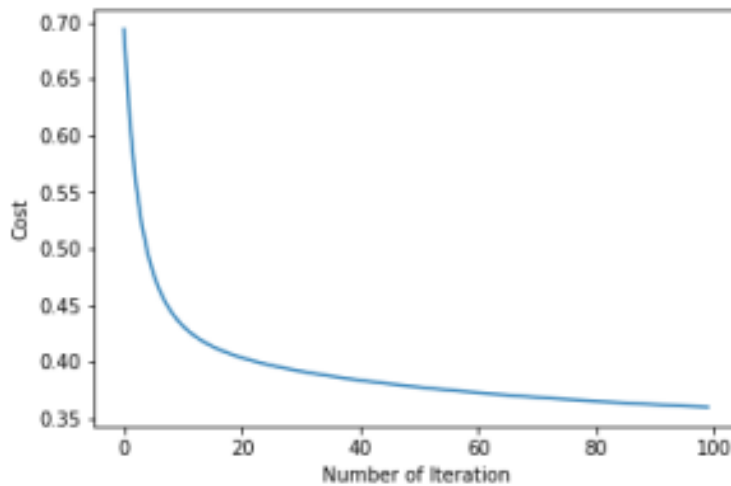


Figure 3: The logistic regression result

And demonstrate the retrieved outcome (Fig. 4). The KNN model showed 88.52% accuracy. We initialize and train the model by the support vectors machine:

```
from sklearn.svm import SVC
svm = SVC(random_state = 1)
svm.fit(x_train.T, y_train.T)
```

Model Accuracy: 85.75%. Naive Bayes method:

```

from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(x_train.T, y_train.T)

```

The model showed an accuracy of 83.19%. Decision trees:

```

from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train.T, y_train.T)

```

The accuracy is 80.33%. We will also use the built-in RandomForest classifier and analyze its result as well:

```

from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators = 1000, random_state = 1)
rf.fit(x_train.T, y_train.T)

```

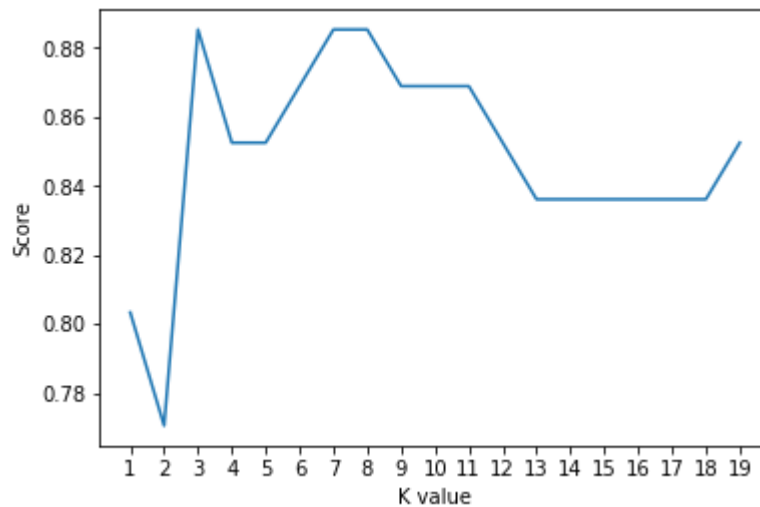


Figure 4: Timeline of changes in the accuracy of the method of K-nearest neighbors.

The accuracy is 88.18%. We initialize all the mentioned models and get the predicted values (Fig. 5):

```

y_head_lr = lr.predict(x_test.T)
knn3 = KNeighborsClassifier(n_neighbors = 3)
knn3.fit(x_train.T, y_train.T)
y_head_knn = knn3.predict(x_test.T)
y_head_svm = svm.predict(x_test.T)
y_head_nb = nb.predict(x_test.T)
y_head_dtc = dtc.predict(x_test.T)
y_head_rf = rf.predict(x_test.T)

```

For a clearer presentation, we create and derive a matrix of errors of each model.

```

from sklearn.metrics import confusion_matrix
cm_lr = confusion_matrix(y_test, y_head_lr)
cm_knn = confusion_matrix(y_test, y_head_knn)
cm_svm = confusion_matrix(y_test, y_head_svm)
cm_nb = confusion_matrix(y_test, y_head_nb)
cm_dtc = confusion_matrix(y_test, y_head_dtc)
cm_rf = confusion_matrix(y_test, y_head_rf)

```

We will output the received result (Fig. 6). Thus, the best result was demonstrated by the model based on the method K of the nearest neighbors. The model showed similar results with the Random

Forest classifier, which is, in fact, a more complex and complex modification of the same method. Therefore, we found that the K-nearest method showed the best result and can be recommended as a priority for use in the expert system.

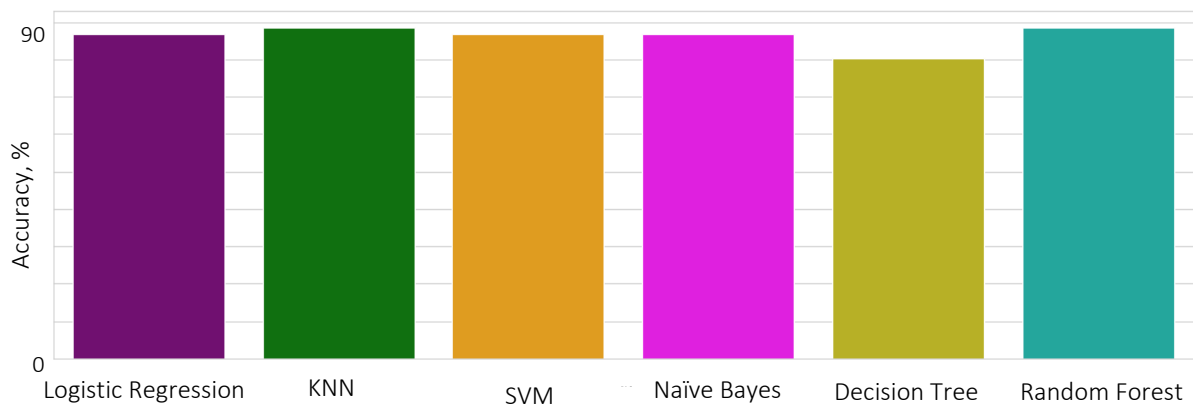


Figure 5: Comparison or estimates of the received models

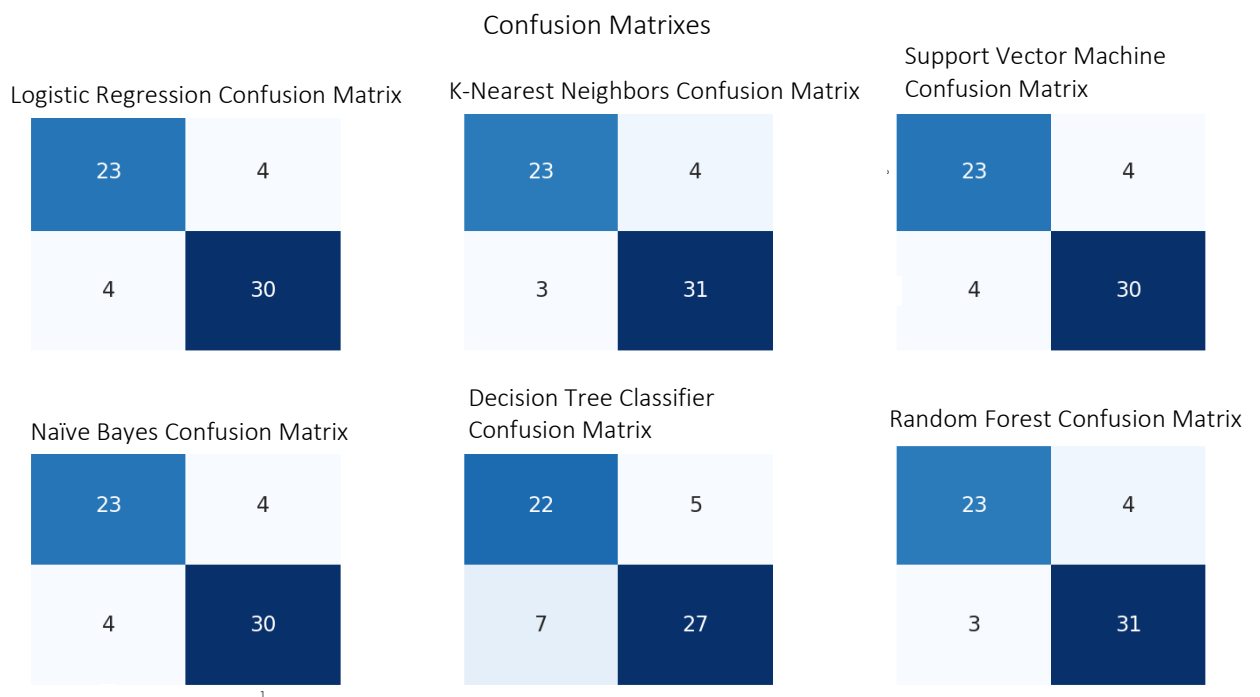


Figure 6: Confusion matrices of the received models.

But now technology is a black box for us - we get the result, but we do not have a clear idea. Therefore we visualize the received tree of decisions:

```

export_graphviz (estimator, out_file='tree.dot',
                 feature_names = feature_names,
                 class_names = y_train_str,
                 rounded = True, proportion = True,
                 label='root',
                 precision = 2, filled = True)

from subprocess import call
call(['dot', '-Tpng', 'tree.dot', '-o', 'tree.png', '-Gdpi=600'])
from IPython.display import Image
Image(filename = 'tree.png')

```

Also, one of the important tools for understanding the model is to establish the influence of attributes on the final result. Below we show how the values of attributes affect the classification of the state.

```
perm = PermutationImportance(model, random_state=1).fit(X_test,
y_test)
eli5.show_weights(perm, feature_names = X_test.columns.tolist())
```

A fragment of the built decision tree show in Fig. 7. Thus, we established which factors most influence the result (Fig. 8). We also visualize not only the relative influence of factors, but also the absolute values that they acquire in a given result. As we see from Fig. 9, we can clearly see how the value of the attribute affects the result and how the values with the same result are grouped. For example, you can see that num_major_vessels directly affects the result, and the values of the attributes are grouped so that the positive values of the result are on the left and the negative values are on the right.

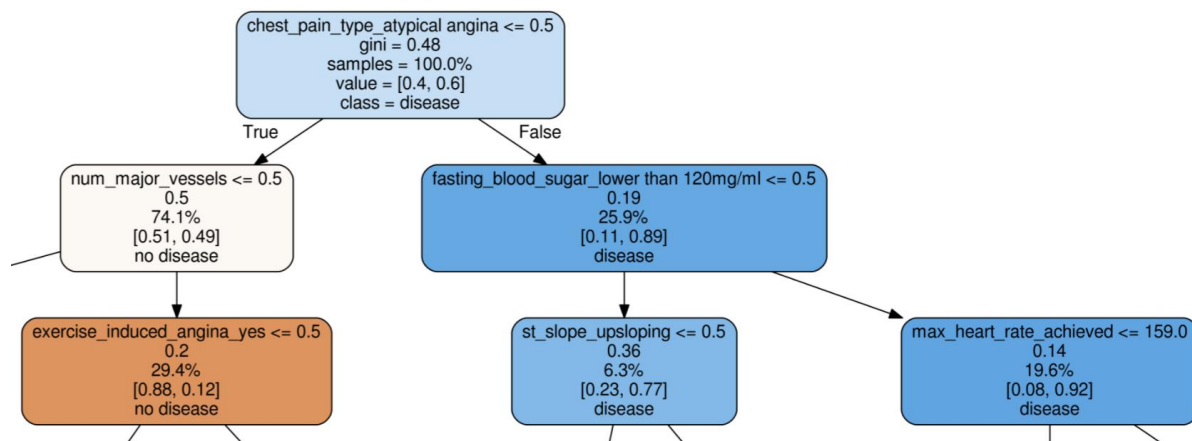


Figure 7: A fragment of the built decision tree

Weight	Feature
0.0754 ± 0.0445	thalassemia_reversable defect
0.0459 ± 0.0321	max_heart_rate_achieved
0.0393 ± 0.0334	num_major_vessels
0.0295 ± 0.0245	st_depression
0.0197 ± 0.0382	thalassemia_fixed defect
0.0197 ± 0.0245	rest_ecg_normal
0.0164 ± 0.0000	exercise_induced_angina_yes
0.0131 ± 0.0131	sex_male
0.0131 ± 0.0131	cholesterol
0.0066 ± 0.0334	age
0 ± 0.0000	chest_pain_type_non-anginal pain
0 ± 0.0000	thalassemia_normal
0 ± 0.0000	fasting_blood_sugar_lower than 120mg/ml
0 ± 0.0000	rest_ecg_left ventricular hypertrophy
0 ± 0.0000	st_slope_upsloping
0 ± 0.0000	resting_blood_pressure
-0.0033 ± 0.0131	chest_pain_type_atypical angina
-0.0131 ± 0.0131	chest_pain_type_typical angina
-0.0328 ± 0.0207	st_slope_flat

Figure 8: Impact of attributes on the received result

3. Conclusion and perspectives

The scope of machine learning applications is constantly expanding. Comprehensive informatization leads to the accumulation of huge amounts of data in science, industry, business, transport, health care. The resulting tasks of forecasting, management and decision-making are often reduced to learning by precedent. Previously, when there was no such data, these tasks were either not set at all, or solved by completely different methods. In this work, machine learning methods for solving practical problems were investigated and implemented. The concept of detection and analysis of heart diseases has been formed, which is proposed to be implemented in the form of an expert system.

The work considered 6 different models for the classification of heart disease. In general, each model has demonstrated high performance and can be used as an auxiliary mechanism for cardiologists and in medical institutions in cardiology departments to diagnose heart disease. The best results were shown by the K-nearest neighbors classifier with a result of 88.52% and the Random Forest classifier built on a similar principle, which is actually directly related to the K-nearest neighbors as its further development.

Other models also showed good results in the range of 80-88%. Among the disadvantages of all systems is the lack of 100% results, and therefore recommend the expert system on their basis as the main tool is not possible. However, each of them can be a very useful addition to the usual, laboratory diagnosis of diseases in the patient. They can be especially useful in combination, which can theoretically improve results. The ensemble (combination) of these models, their integration with each other can be the further development of the system implemented in the work.

One of the important tasks was not only to build a machine learning model, but also to explain how it works, what impacts the result and how it is calculated. Therefore, an algorithm for examining the obtained result with the analysis and research of the impact of attributes and their visualization was presented, which allowed us to investigate the received result in more detail and transparency.

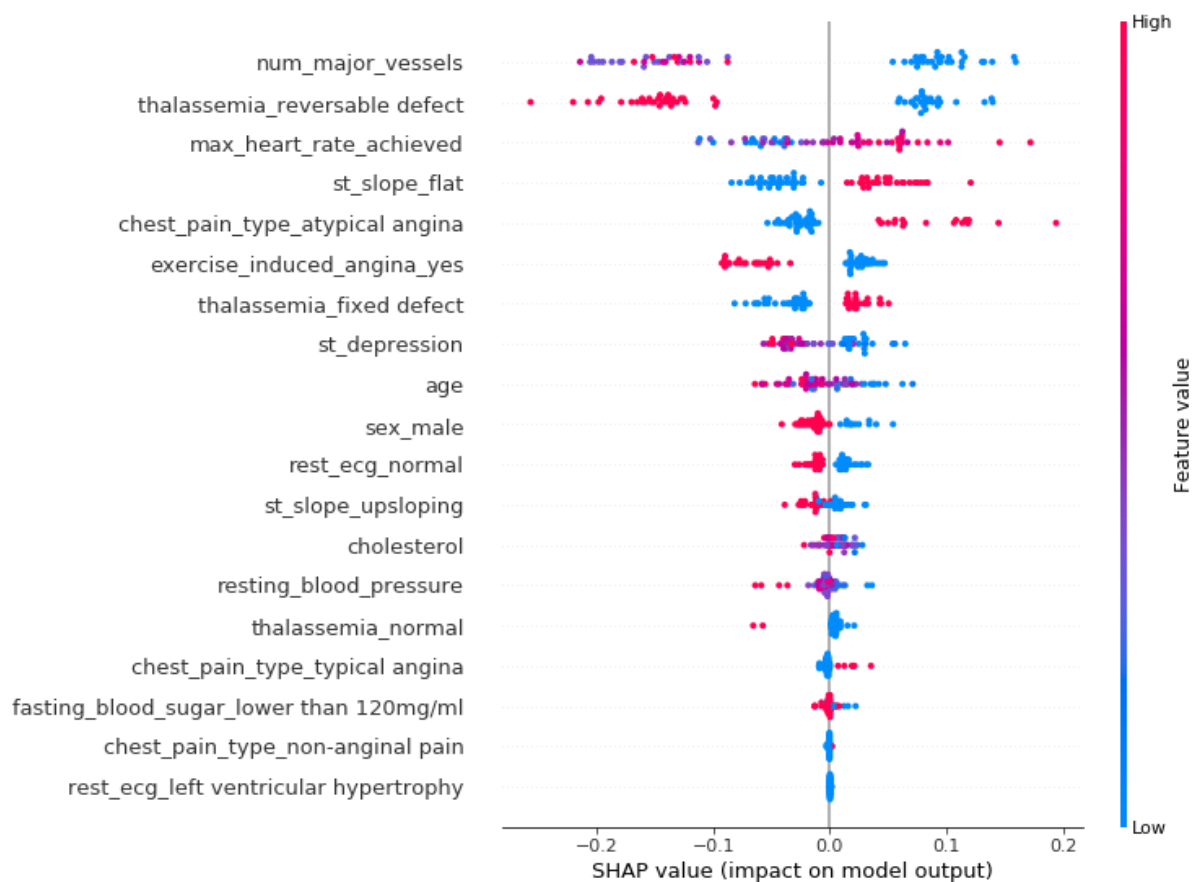


Figure 9: Distribution of attribute values depending on the result

4. References

- [1] Institute for Health Metrics and Evaluation. Global Burden of Disease (GBD 2019). 9 Dec. 2020. URL: <http://www.healthdata.org/gbd/2019>.
- [2] Krucevych, T. Henderni. Osoblyvosti proyavu faktoriv ryzyku sercevo-sudynnyx zaxvoryuvan u cholovikiv i zhinok zriloho viku. Sportyvnyj visnyk Prydniprov'ya. 2019. № 3. S. 110-117. (In Ukrainian).
- [3] Daniel M. Study Suggests Medical Errors Now Third Leading Cause of Death in the U.S. [Internet resource]. Mode of access: https://www.hopkinsmedicine.org/news/media/releases/study_suggests_medical_errors_now_third_leading_cause_of_death_in_the_us
- [4] Yevropejs"ke rehional"ne byuro VOOZ. Doslidzhennya STEPS: poshyrenist" faktoriv ryzyku neinfekcijnyx zaxvoryuvan" v Ukrayini u 2019 roci. Kopenhahen, 2020. License: CC BY-NC-SA 3.0 IGO. (In Ukrainian).
- [5] Koktysh L. The state of the art in health data analytics [Internet resource]. Mode of access: <https://www.scnsoft.com/blog/health-data-analytics-overview>
- [6] Ty - Jourti. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms VL - 2018PY - 2018DA - 2018/12/02DO - 10.1155/2018/3860146 UR, [online]. URL: <https://doi.org/10.1155/2018/3860146AB>.
- [7] Rashmee U. Shah and John S. Rumsfeld, Big Data in Cardiology European Heart Journal, vol. 38, no. 24, pp. 1865-1867, June 2017, [online]. URL: <https://doi.org/10.1093/eurheartj/ehx284>.
- [8] IBM Watson Health [Internet resource]. URL: <https://www.ibm.com/watson/health/oncology-and-genomics/oncology/>
- [9] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.* 2018;19(6):1236-1246. doi:10.1093/bib/bbx044
- [10] Fan J, Xing L, Ma M, Hu W, Yang Y: Verification of the machine delivery parameters of treatment plan via deep learning. *Physics in Medicine and Biology, Early Access*, 2020.
- [11] Top 6 use cases of data science in healthcare, Jan. 2019, [online]. URL: <https://bigdata-madesimple.com/top-6-use-cases-of-data-science-in-healthcare/>.
- [12] Cheryl Alexander and Lidong. Wang, Big Data Analytics in Heart Attack Prediction. *Journal of Nursing & Care.* 06. 10.4172/2167-1168.1000393, 2017.
- [13] Khlevna I., Koval B. Fraud detection technology in payment systems. // IT&I 2020 – Information Technology and Interactions. Proceedings of the 7th International Conference "Information Technology and Interactions" (IT&I-2020). Workshops Proceedings. Kyiv, Ukraine, December 02-03, 2020. CEUR Workshop Proceedings, – P. 85 – 95.
- [14] Kalra D., Beale T., Heard S., Lloyd D. An HER architecture for Archetyped Health Information Systems, 2003. URL: <http://www.openehr.org>.
- [15] Bresnick J. How to Choose the Right Healthcare Big Data Analytics Tools [Internet resource]. URL: <https://healthitanalytics.com/features/how-to-choose-the-right-healthcarebig-data-analytics-tools>
- [16] Heart diseases dataset [Internet resource]. URL: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>.
- [17] Teslia I., Yehorchenkov O., Khlevna I., Khlevnyi A. Development concept and method of formation of specific project management methodologies. 2018. *Eastern European Journal of Advanced Technologies*, 5/3(95), 6 – 16.