# Analysis of Machine Learning Algorithms for Classification and Prediction of Heart Disease

Nataliya Boyko and Iryna Dosiak

*Lviv Polytechnic National University, Profesorska Street 1, Lviv, 79013, Ukraine*

**Abstract**

The study aims to improve the effectiveness of health care in various ways. The paper considers ML algorithms that allow health professionals to allocate resources optimally and physicians to choose the best treatment options for patients. This approach will reduce the burden on doctors and increase and accelerate patients' access to health care, save resources and reduce costs. The paper presents the results of research that will allow the use of smaller data sets to develop transparent models. The report uses a naive Bayes classifier to predict heart disease. The advantage of this approach is that the sample size requirements are reduced from exponential to linear, which is very important. There is an overview of the classification model, its advantages and disadvantages. Materials and methods are also analyzed.

**Keywords 1**

Model, classification, machine learning, algorithm, Bayes classifier

## 1. Introduction

Machine Learning (ML) algorithms allow healthcare professionals to allocate resources optimally and physicians to choose the best treatment options for patients. This approach reduces the burden on doctors, increases and accelerates patients' access to health care, saves resources, and reduces costs. However, despite the achievements of ML research in medicine, its role is currently limited. Creating and testing a model may require large amounts of high-quality data. Besides, diagnostic models must be built individually for each disease. It is a lengthy process. In addition, the psychological aspect of trusting black box algorithms can also be difficult to perceive. However, continuing ML research may allow using smaller data sets and developing more transparent models [4, 13].

The nature of heart disease is complex. In addition, the diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data. The relationship between the real cause of the disorder and the effects of spontaneous symptoms in patients can often be hidden and not obvious [6].

That is why the analysis of medical data in health care is considered an important but complex task that must be performed accurately and effectively. In addition, the study of medical data is necessary to avoid medical error.

The basis of medical diagnosis is the problem of classification. The diagnosis comes down to the problem of displaying data to one of N different results.

The study aims to apply and implement the original Naive Bayes model with two existing models: the Gaussian model and the Multinomial model.

This study will focus on comparative analysis, differences, capabilities, and effectiveness of the classifier with different models

IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine, November 19–21, 2021 Valencia, Spain

EMAIL: nataliya.i.boyko@lpnu.ua (N. Boyko); iryna.dosiak.knm.2018@lpnu.ua(I. Dosiak)
ORCID: 0000-0002-6962-9363 (N. Boyko); 0000-0002-5488-4468 (I. Dosiak)

The purpose of classifying heart disease is to diagnose a disease in a patient based on specific diagnostic measurements included in the data set. In addition, the work will consist of searching for significant features and patterns between the various factors influencing the diagnosis.

## 2. Review of literature sources

For a detailed study of these tasks, you need to read and analyze the experience of scientists in this field. Since the problem is relevant, numerous studies have been conducted that have focused on diagnosing heart disease in combination with or without another condition.

- G. Parthiban, A. Rajesh, S.K. Srivatsa predicted the chances of people with diabetes having heart disease and highlighted the results in their article "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method," published in the International Journal of Computer Applications [1 ]. The accuracy was 74%.
- Mrs. Mr. Subbalakshmi, Mr. K. Ramesh M. Tech, Mr. M. Chinna Rao M.Tech developed a system that extracts hidden knowledge from a historical heart disease database using a Naive Bayes classification [2]. The article "Decision Support in Heart Disease Prediction System using Naive Bayes" was published in the Indian Journal of Computer Science and Engineering».
- Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni conducted a study and compared KNN and the Naive Bayes classifier to predict heart disease [3]. However, the accuracy of the results reached 45.6% for KNN and 52.33% in the case of the Naive Bayes classifier. Their article "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" was published in the International Journal of Computer Applications. In the end, they added the need to improve the proposed study.
- Vincy Cherian and Bindu M.S developed a heart disease prediction system using a Naive Bayes classifier and a Laplace smoothing technique [4]. They reported this in their article "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques." They achieved high accuracy. However, the system has a limit on the number of attributes - symptoms.

Unfortunately, searches for such studies among Ukrainian sources did not yield any results.

Thus, various studies only represent the effectiveness of predicting heart disease using ML methods. This study aims to find features and patterns between different factors that affect the diagnosis using a Naive Bayes classifier.

## 3. Methods overview

Classification solves the following problem: let there be a set of objects divided into classes on one or more grounds. Moreover, a finite set of objects is given, for which it is known to which classes they belong. Such a set is considered to be a training sample. It is unknown to which class the other objects belong. We need to build an algorithm that can classify any object of the source set - specify the number or name of the class to which it belongs [9, 11].

### 3.1. A mathematical formulation of the classification problem

Let *X* be a set of object descriptions, and *Y* be class numbers or names. There is an unknown target relationship - mapping $y^*: X \rightarrow Y$, the values of which are known only on the elements of the finished training sample $X^m = \{(x_1, y_1), \dots (x_m, y_m)\}$. We need to build an algorithm $a: X \rightarrow Y$, that can classify an arbitrary object $x \in X$ [12].

### 3.2. Bayes classifier

Bayes classifier - provides a classification with a degree of confidence rather than simply issuing the most plausible class. Bayes' theorem is used to determine the degree of certainty.

Bayes' theorem describes the probability of an event, given the circumstances that may affect the event. Thus, you can more accurately calculate the probability, considering both already known information and data from new observations [14].

A Naive Bayes classifier is an assumption about the independence of traits. In other words, the NCB assumes that any attribute in the class is not related to the presence of any other feature.

## 3.3. Method overview

As mentioned, the Bayes classifier is based on the Bayes theorem, which describes the probability of an event, given the circumstances that may affect the event [14].

Suppose there is a symptom *S*. In addition, there are classes (diseases) *C*, which should include the symptom. It is necessary to find a class (disease) *C* in which the probability for this line would be maximum. The mathematical notation is given in Formula 1.

$$c = \underset{C}{argmax}\, P(C|S) \tag{1}$$

It is hard to calculate *P(C|O)*. However, you can use Bayes' theorem and go to (Formula 2):

$$P(C|O) = \frac{P(S|C)\,P(C)}{P(S)}, \tag{2}$$

where *P(C)* - an a priori probability, the probability of meeting a class among all the data;

*P(O|C)* - conditional probability, the probability of symptoms in each class;

*P(O)* - total probability, probability of symptoms.

Usually, it makes no sense to work with one symptom. It is much more effective to detect the disease on several grounds. Thus Formula 2 will take the form (Formula 3):

$$P(C|S_1S_2\dots S_n) = \frac{P(S_1S_2\dots S_n|C)P(C)}{P(S_1S_2\dots S_n)} \tag{3}$$

Since you need to find the function's maximum, the denominator can be ignored (this is a constant). It is also necessary to include a "naive" assumption that the symptoms of *S* depend only on class *C* and do not depend on each other. Then the numerator will take the form (Formula 4):

$$P(C)P(S_1|C)P(S_2|C_{S_1})\dots P(S_n|C_{S_1S_2\dots S_n}) = P(C)P(S_1|C)P(S_2|C)\dots P(S_n|C) = P(C)\prod_i(S_i|C) \tag{4}$$

So, the final formula will look like (Formula 5):

$$c = \underset{C}{argmax}\, P(C)\prod_i(S_i|C) \tag{5}$$

So it all comes down to calculating the probability *P(C)* and *P(S|C)*. Calculating these parameters is called classifier training.

## 3.4. Multinomial Naive Bayes

Multinomial Naive Bayes implements a Naive Bayes algorithm for multinomial distributed data and is one of two classic variants of Naive Bayes [8].

This algorithm puts forward a second assumption of independence - the assumption of positional independence. Conditional probabilities of symptom onset are equally independent of its position in the data sample [9].

The data is usually presented as a vector. The basic idea is that each unique feature (symptom) that occurs is assigned a unique integer. Therefore the data can be represented as a sequence of numbers.

The distribution of the number of vectors is parameterized by vectors $\theta_c = (\theta_{c_1} \dots \theta_{c_n})$ for each class, where *n* - number of features (symptom), and $\theta_{c_1}$ – the probability $(S_i|C)$ of the appearance in the sample of features belonging to class *C*.

The parameter $\theta_c$ is estimated by the smoothed version of the maximum probability. The relative frequency calculation (Formula 6):

$$\theta_{c_i} = \frac{N_{c_i} + a}{N_c + a_n}, \tag{6}$$

where $N_{c_i}$ - the number of times the *i* character appears in a class *C* sample in the training set.;

$N_c$ - the total number of all features (symptoms) for class *C*;
*A* - Laplace smoothing.

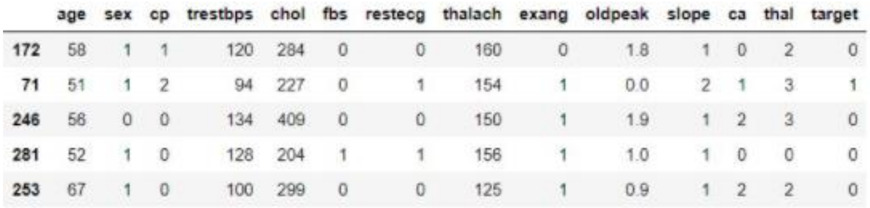
## 4. Review and analysis of data

The data set about heart disease "heart.csv" is used for research [6]. It was taken from Kaggle. This database contains 76 attributes, but all published experiments involve using a subset of 14 of them, as the rest of the information is the identification of individuals. The total number is 303 rows and 14 columns, of which 165 have heart disease [7].

Attribute information:

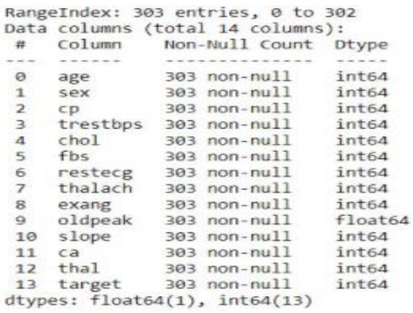
1. age;
2. sex : (1 = a man; 0 = a woman);
3. cp: chest pain type (4 values);
4. trestbps: blood pressure at rest (in mm Hg on admission to the hospital);
5. chol: serum cholesterol in mg / dl;
6. fbs: (fasting blood sugar) (1 => 120 mg / dl; 0 = <120 mg / dl);
7. restecg: the results of electrocardiography at rest (values 0, 1, 2);
8. thalach: the maximum pulse;
9. exang: angina caused by exercise (1 = yes; 0 = no);
10. oldpeak: ST depression caused by exercise for rest;
11. slope: the slope of the peak segment of exercise ST;
12. ca: the number of major vessels (0–3) stained by fluoroscopy;
13. thal: thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect);
14. target: (1 = heart disease or 0 = no heart disease).

Fig. 1, Fig. 2, and Fig. 3 show a dataset.



| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **172** | 58 | 1 | 1 | 120 | 284 | 0 | 0 | 160 | 0 | 1.8 | 1 | 0 | 2 | 0 |
| **71** | 51 | 1 | 2 | 94 | 227 | 0 | 1 | 154 | 1 | 0.0 | 2 | 1 | 3 | 1 |
| **246** | 56 | 0 | 0 | 134 | 409 | 0 | 0 | 150 | 1 | 1.9 | 1 | 2 | 3 | 0 |
| **281** | 52 | 1 | 0 | 128 | 204 | 1 | 1 | 156 | 1 | 1.0 | 1 | 0 | 0 | 0 |
| **253** | 67 | 1 | 0 | 100 | 299 | 0 | 0 | 125 | 1 | 0.9 | 1 | 2 | 2 | 0 |

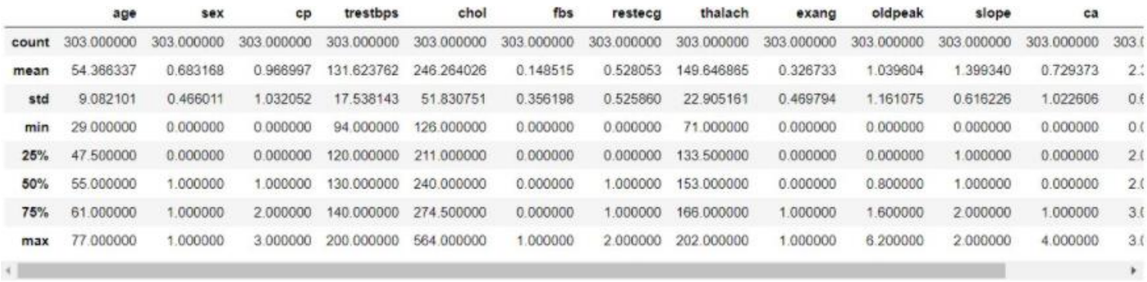
**Figure 1:** Image of the first five rows of data



```
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
```

**Figure 2**: Attributes overview

From Fig. 2, we can see that the categorical data are missing. There are numeric data of type int and float.

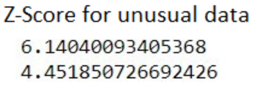


| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.( |
| **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.: |
| **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.( |
| **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.( |
| **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.( |
| **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.( |
| **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.( |
| **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.( |

**Figure 3**: The main data characteristics

As can be seen from this section, most values are usually categorized. All columns have no spaces, contain 303 rows of data.

An analysis of atypical emissions should also be conducted. To do this, use a standardized Z-Score score, which shows how many standard deviations is the scatter of the value relative to the observed average value. If the Z-Score value is greater than or less than 3 or -3, respectively, this data point will be defined as non-standard (Fig. 4).



```
Z-Score for unusual data
 6.14040093405368
 4.451850726692426
```

**Figure 4**: Z-Score for atypical data

Fig. 5 shows that this data set contains two emissions. Let's try to visualize them. For this purpose, it is necessary to construct the box diagram to visualize atypical data (Fig. 5).

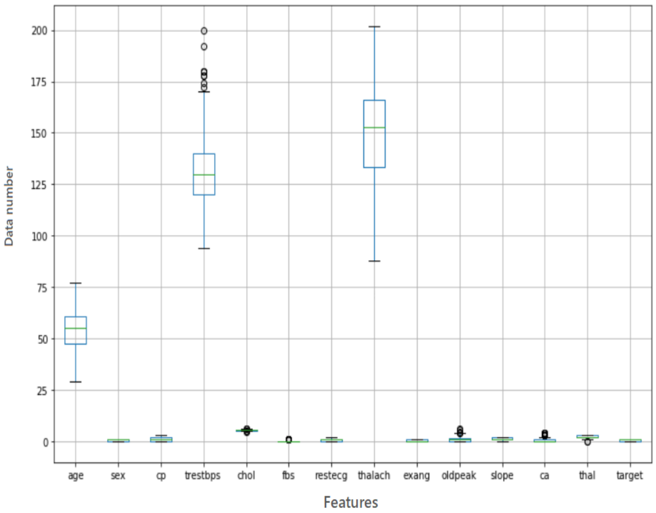
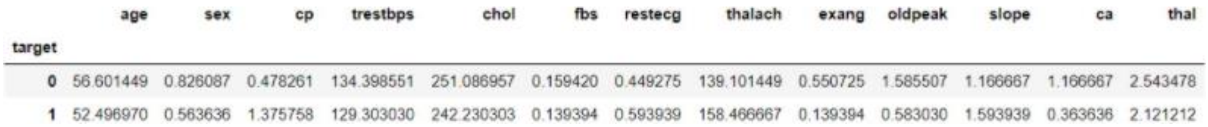


**Figure 5**: Visualization of atypical data in a dataset

Because only two sets of data that differed from the others were identified, so they were removed from the sample. This will help achieve better results in predicting heart disease.

The next step is to review the number of existing or absent diseases. To do this, determine the average number of different values for prediction by columns (Fig. 6).



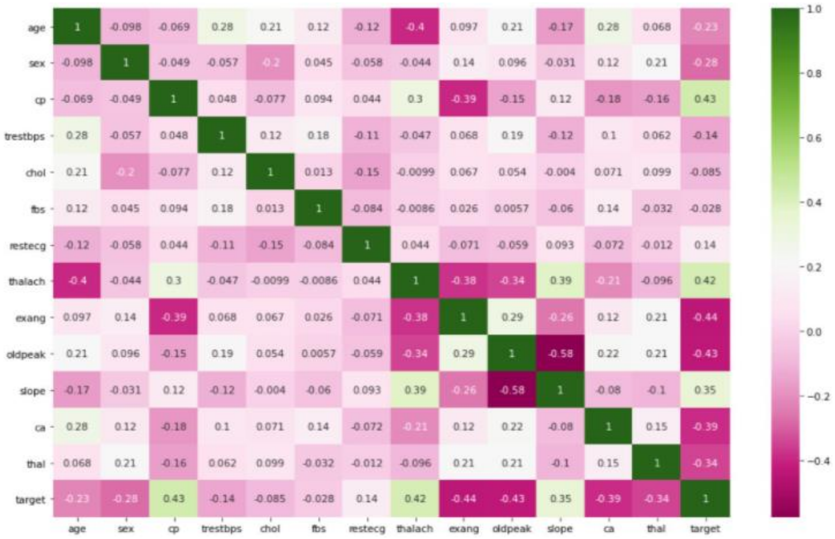
| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **target** | | | | | | | | | | | | | |
| **0** | 56.601449 | 0.826087 | 0.478261 | 134.398551 | 251.086957 | 0.159420 | 0.449275 | 139.101449 | 0.550725 | 1.585507 | 1.166667 | 1.166667 | 2.543478 |
| **1** | 52.496970 | 0.563636 | 1.375758 | 129.303030 | 242.230303 | 0.139394 | 0.593939 | 158.466667 | 0.139394 | 0.583030 | 1.593939 | 0.363636 | 2.121212 |

**Figure 6**: An image of the mean values for the column, which determines the presence or absence of the disease

Target variable: whether the patient has heart disease or not (value 0 - yes; value 1 - no). Fig. 6 shows that the distribution is balanced.

## 4.1. Search for the correlation of heart disease with different parameters

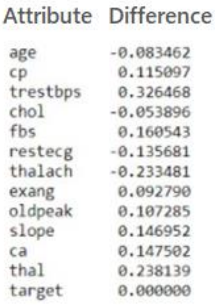
To find the links of heart disease with different parameters, we need to build a correlation matrix (Fig. 7).



| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1 | -0.098 | -0.069 | 0.28 | 0.21 | 0.12 | -0.12 | -0.4 | 0.097 | 0.21 | -0.17 | 0.28 | 0.068 | -0.23 |
| sex | -0.098 | 1 | -0.049 | -0.057 | -0.2 | 0.045 | -0.058 | -0.044 | 0.14 | 0.096 | -0.031 | 0.12 | 0.21 | -0.28 |
| cp | -0.069 | -0.049 | 1 | 0.048 | -0.077 | 0.094 | 0.044 | 0.3 | -0.39 | -0.15 | 0.12 | -0.18 | -0.16 | 0.43 |
| trestbps | 0.28 | -0.057 | 0.048 | 1 | 0.12 | 0.18 | -0.11 | -0.047 | 0.068 | 0.19 | -0.12 | 0.1 | 0.062 | -0.14 |
| chol | 0.21 | -0.2 | -0.077 | 0.12 | 1 | 0.013 | -0.15 | -0.0099 | 0.067 | 0.054 | -0.004 | 0.071 | 0.099 | -0.085 |
| fbs | 0.12 | 0.045 | 0.094 | 0.18 | 0.013 | 1 | -0.084 | -0.0086 | 0.026 | 0.0057 | -0.06 | 0.14 | -0.032 | -0.028 |
| restecg | -0.12 | -0.058 | 0.044 | -0.11 | -0.15 | -0.084 | 1 | 0.044 | -0.071 | -0.059 | 0.093 | -0.072 | -0.012 | 0.14 |
| thalach | -0.4 | -0.044 | 0.3 | -0.047 | -0.0099 | -0.0086 | 0.044 | 1 | -0.38 | -0.34 | 0.39 | -0.21 | -0.096 | 0.42 |
| exang | 0.097 | 0.14 | -0.39 | 0.068 | 0.067 | 0.026 | -0.071 | -0.38 | 1 | 0.29 | -0.26 | 0.12 | 0.21 | -0.44 |
| oldpeak | 0.21 | 0.096 | -0.15 | 0.19 | 0.054 | 0.0057 | -0.059 | -0.34 | 0.29 | 1 | -0.58 | 0.22 | 0.21 | -0.43 |
| slope | -0.17 | -0.031 | 0.12 | -0.12 | -0.004 | -0.06 | 0.093 | 0.39 | -0.26 | -0.58 | 1 | -0.08 | -0.1 | 0.35 |
| ca | 0.28 | 0.12 | -0.18 | 0.1 | 0.071 | 0.14 | -0.072 | -0.21 | 0.12 | 0.22 | -0.08 | 1 | 0.15 | -0.39 |
| thal | 0.068 | 0.21 | -0.16 | 0.062 | 0.099 | -0.032 | -0.012 | -0.096 | 0.21 | 0.21 | -0.1 | 0.15 | 1 | -0.34 |
| target | -0.23 | -0.28 | 0.43 | -0.14 | -0.085 | -0.028 | 0.14 | 0.42 | -0.44 | -0.43 | 0.35 | -0.39 | -0.34 | 1 |

**Figure7**: Correlation matrix

Fig. 7 shows certain relationships between the features. It is first necessary to determine the difference between the correlation coefficients in men and women. The results are shown in Fig. 8.



```
Attribute  Difference

age        -0.083462
cp          0.115097
trestbps    0.326468
chol       -0.053896
fbs         0.160543
restecg    -0.135681
thalach    -0.233481
exang       0.092790
oldpeak     0.107285
slope       0.146952
ca          0.147502
thal        0.238139
target      0.000000
```

**Figure 8**: Difference of correlation coefficients for different sexes

Figure 8 shows that all coefficients, except for the target variable, differ between men and women. The most noticeable difference for trestbps. This is the resting blood pressure in millimeters of mercury.

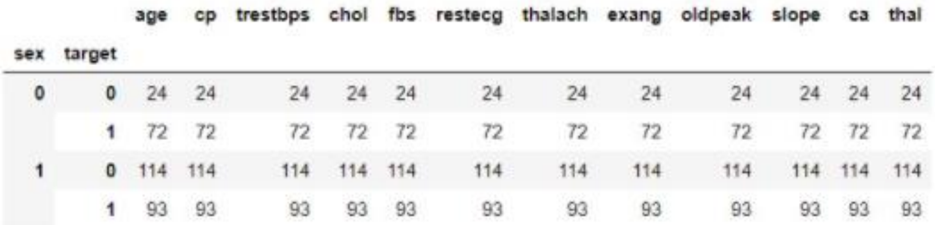
Most people have normal blood pressure in certain groups (these can be healthy adults, adults taking medication, the elderly). It also appears that very high blood pressure may indicate heart disease [14, 15].

Observations follow from the obtained results:

1. Age is negatively correlated with heart disease. Because older people are more likely to have heart disease, they are more likely to have a health check-up, even if they have mild or no symptoms. Young people go for a health check only when they have apparent symptoms. That is why they are more often diagnosed with heart disease.
2. Cholesterol and fasting blood glucose levels have little correlation with heart disease.
3. Chest pain (cp), maximal pulse (thalach), a tilt of the ST segment in the ECG are positively correlated with heart disease.
4. Exercise angina (exang), ST depression caused by exercise (oldpeak), the number of major vessels (0-3) stained with fluoroscopy (ca) are negatively correlated with heart disease. Moreover, in all these ratios, the correlation is lower for men than for women.

5. Trestbps (resting blood pressure) and fbs (fasting blood sugar) are negatively correlated. Moreover, the correlation is lower for women compared to men.

For these observations, the accuracy of the conclusions should be checked, taking into account the distribution of data between men and women (Fig. 9).



| sex | target | age | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 | 24 |
|  | 1 | 72 | 72 | 72 | 72 | 72 | 72 | 72 | 72 | 72 | 72 | 72 | 72 |
| 1 | 0 | 114 | 114 | 114 | 114 | 114 | 114 | 114 | 114 | 114 | 114 | 114 | 114 |
|  | 1 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 |

**Figure 9**: Data distribution between women and men

Fig. 9 shows that women account for about half of the observations than men. You can also see that gender is a risk factor.

Also, to verify the above statements, you should visualize the presence or absence of the disease depending on the age range (Fig. 10).

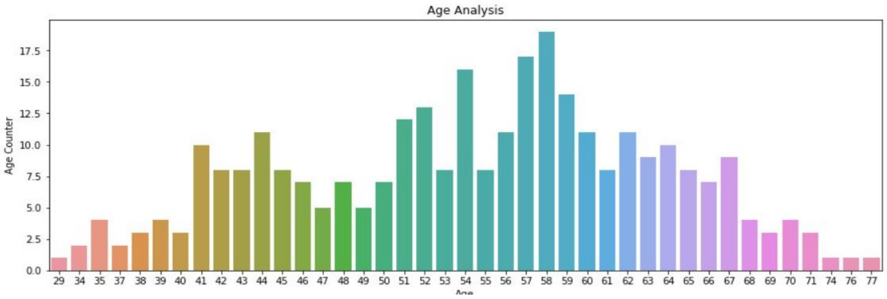


**Figure 10**: The amount of data corresponding to each age

In Fig. 10 x-axis indicates the age of patients, y-axis - the number of patients of a certain age. The graph shows that the age of the youngest patient is 22, the oldest is 77. The most common patients are aged 58. There are few patients under 40 or after 70. Therefore, the age distribution is shown in Fig. 11.

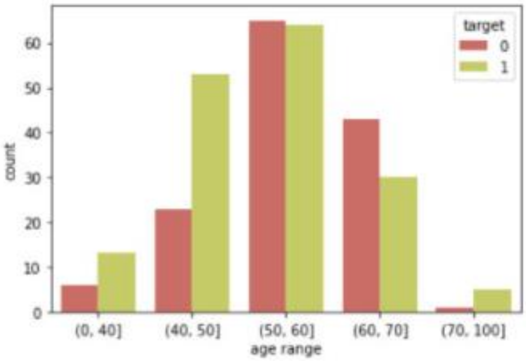


**Figure 11**: The presence or absence of the disease in different age categories

Fig. 11 shows the age distribution: from 0 to 40 years, from 40 to 50, from 50 to 60, from 60 to 70, from 70 to 100. Green shows the presence of the disease, red the absence. The age range is arranged along the x-axis and the number of patients along the y-axis.

In Fig. 12, the x-axis represents the age of the patients. The y-axis represents the density estimate. Yellow indicates the absence of the disease, red - the presence. The relationship between age and female gender on the left. Between age and male - on the right.

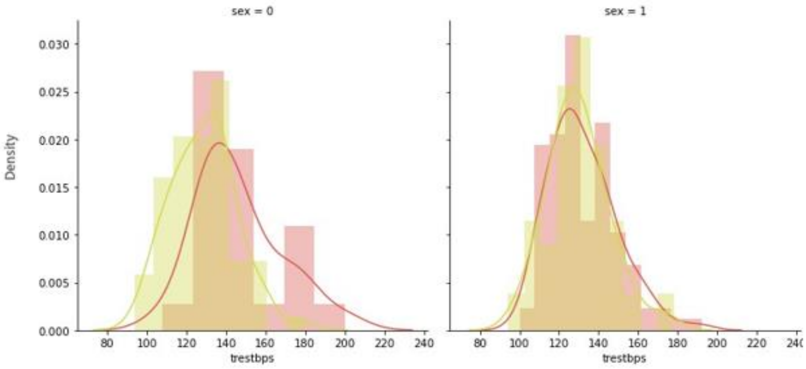


**Figure 12**: Relationship between blood pressure and sex

In Fig. 12, the x-axis represents the resting blood pressure in millimeters of mercury. The y-axis represents the density estimate. Yellow indicates the absence of the disease, red - the presence. The relationship between blood pressure and the female sex is on the left, on the right - between blood pressure and male.

In Fig. 13. presents the presence or absence of the disease, taking into account only one feature - one attribute.

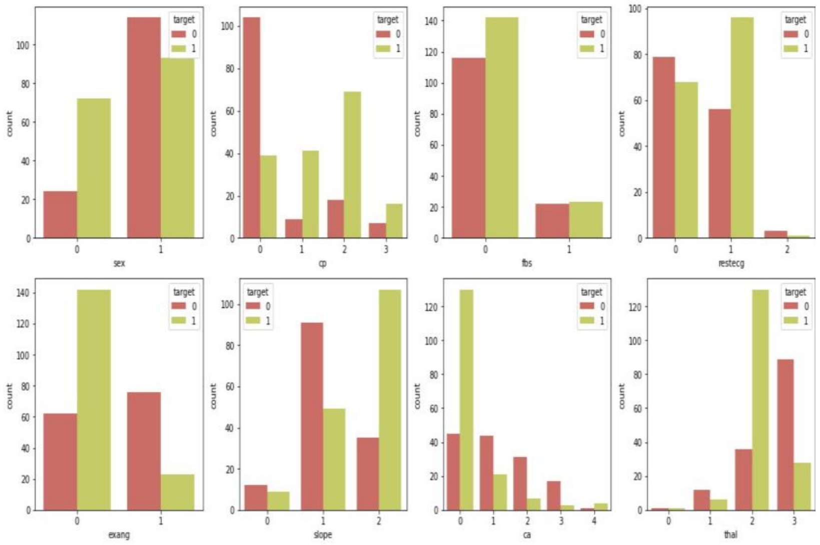


**Figure 13**: Relationship between the presence of the disease and other attributes

In Fig. 13, x-axis shows features: gender, chest pain, blood sugar, electrocardiogram results, angina, ST-segment tilt during the most difficult part of the exercise, the number of major vessels stained with fluoroscopy, and thalassemia. The y-axis shows the number of patients. Yellow indicates the presence of the disease, red - the absence.

From Fig. 13, the following observations follow:

6. The number of major vessels stained with fluoroscopy refers to the number of narrow vessels seen, so the higher the value of this feature, the greater the likelihood of heart disease.
7. A very invasive process for patients obtains the results of blood flow observed through the radioactive dye. But in themselves, they are excellent evidence of heart disease or not.
8. The slope of the ST segment can help determine if you have heart disease or not if it is flat or growing.
9. Angina is a good indicator of heart disease. However, we can also see that knowing what angina is and what it is not an easy task can be confused with other pains or atypical angina.
10. When someone has heart disease, the first symptom is usually stable angina (angina during exercise). When angina occurs even at rest, the condition worsens (typically narrowing the coronary arteries). That is why so few patients find abnormal heart rates at rest, and the vision of this anomaly is very indicative of the presence of heart disease.

11. On the other hand, a value of 0, the probable presence of hypertrophy, in itself does not indicate the presence of heart disease.
12. In itself, the feature - blood sugar levels, does not give confidence in the presence or absence of heart disease. However, we will not abandon this feature, as it can be helpful with other variables.
13. Chest pain also does not give an unambiguous answer. It is challenging to tell if a patient has heart disease that corresponds only to its symptoms.

To verify the accuracy of the conclusions, you should use PCA, which helps extract a set of variables from an existing large set of variables. These extracted variables are called essential components.

Because the data set is small and has no many features, only two components should be used to see how much variance it covers.

The study can explain approximately 90% of the variance in the data set using only two components. Fig. 14 presents each of these decomposed components:

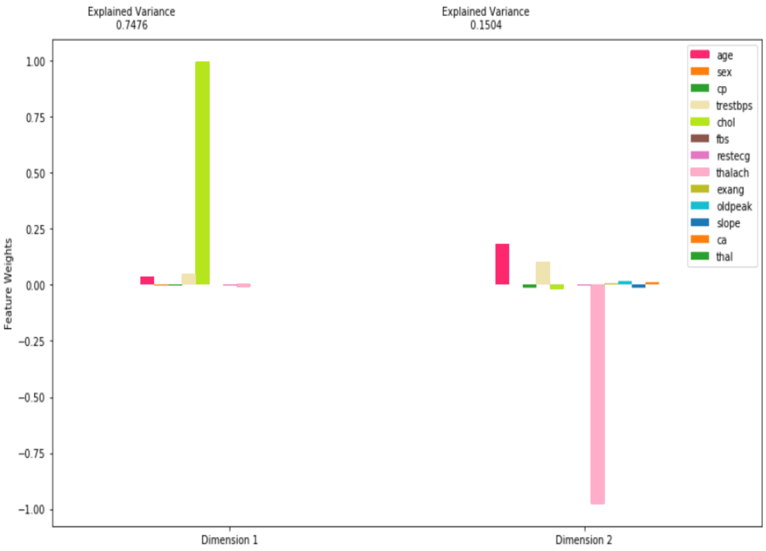


**Figure 14**: Analysis of the main components

Component 1:

Fig. 14 shows that the weight is considerable and positive for the feature of chol, slightly positive for sex and cp. This means that patients with a high rate of this component will have a meager chance of being diagnosed with heart disease. At the same time, people with more elevated serum cholesterol are more likely to be diagnosed with heart disease.

Component 2:

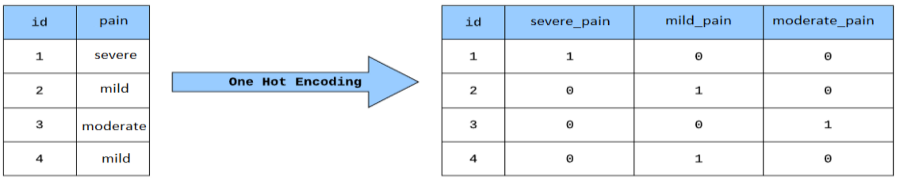
Fig. 14 shows that the weight is considerable and negative for thalach (maximum heart rate reached) and slightly negative for cp (type of chest pain), chol (serum cholesterol), and slope (slope of the peak segment of exercise ST).

Thus, a high rate of thalach, cp, slope and chol, mainly does not cause heart disease. People who have high levels of these components are much less likely to have heart disease. In contrast, age and high resting blood pressure (trestbps) may be the first features of heart disease. In Fig. 14, they are positive.

## 4.2. Application of the Naive Bayes classifier

The next step is to divide the data into training and test in 80% to 20%. You should also normalize the data with OneHotEncoder and MinMaxScaler [10].

OneHotEncoder - a strategy in which each value of the category is converted into a new column, and it is assigned a value of 1 or 0 (notation for true/false). Fig. 15 shows an example of the strategy.



| id | pain |
|---|---|
| 1 | severe |
| 2 | mild |
| 3 | moderate |
| 4 | mild |

| id | severe_pain | mild_pain | moderate_pain |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 1 | 0 |

**Figure 15**: Example of OneHotEncoder operation

Fig. 15 shows an example of the OneHotEncoder operation. The pain column, which contained three classes: medium, strong, and weak, was divided into three new columns: severe pain, moderate pain, and mild pain. All columns contain only two values: 1 if the information is confirmed, 0 if not.

For each value in the object, MinMaxScaler subtracts the minimum value and then divides it by range. The range is the difference between the initial maximum and the initial minimum. MinMaxScaler retains the shape of the original distribution.

After normalization, the classification should be performed. To implement the classification, you need to use GuassianNB from the sklearn library with different types of states when sharing data.

The score function from the sklearn library is used to evaluate the results, which returns the average accuracy of the given test data and labels. The results obtained are presented in Fig.16.

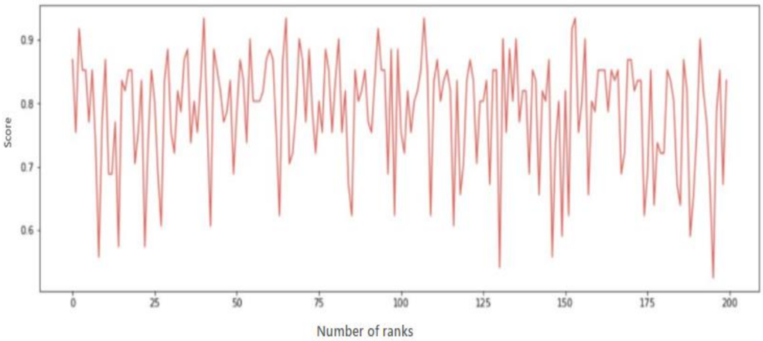


**Figure 16**: Average classification scores using all attributes

Fig. 16 x-axis indicates the number of random states, y-axis - the average score for this method. Fig. 16 shows that the estimate ranges from 0.5 to 1.

Thus, the average estimate of the Naive Bayes classifier for random states from 0 to 200 is 0.844262295081968.

To illustrate the performance of the algorithm should build a matrix of inconsistencies (confusion matrix).

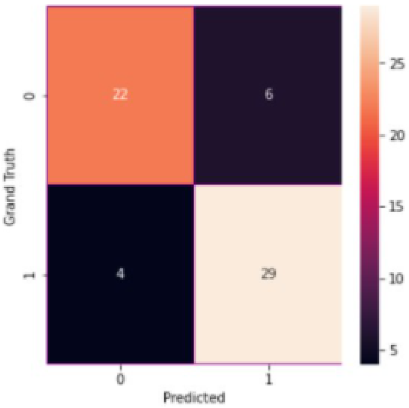


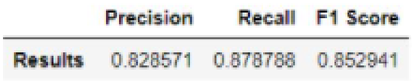
**Figure 17**: Matrix of discrepancies

Figure 17 shows four different results: true positive, false positive, true negative, and false negative.

From the correlation matrix, you can determine the accuracy or positive predictive value (precision), the probability of detection (recall), and the completeness of the definition (f1_score).

- TP - true-positive decision;
- TN - true-negative decision;
- FP - false-positive decision;
- FN - false-negative decision.

The next step is to use the metrics for this method. The results are shown in Fig. 18.



|  | Precision | Recall | F1 Score |
|---|---|---|---|
| Results | 0.828571 | 0.878788 | 0.852941 |

**Figure 18**: Measures of accuracy

We need to reduce the number of attributes to 10. To do this, we need to remove the parameters that have the most negligible impact on heart disease and apply the Naive Bayes classifier again. The results of the experiment are shown in Fig. 19.

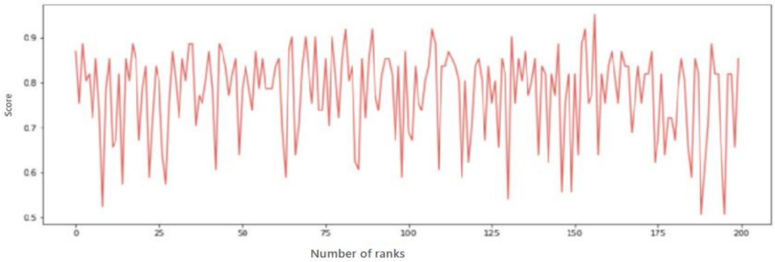


**Figure 19**: Average classification scores using 10 attributes

Fig. 19 x-axis indicates the number of random states, y-axis - the average score for this method. The figure shows that the score ranges from 0.4 to 1.

Thus, the average estimate of the naive Bayes classifier for random states from 0 to 200 is 0.830327868852459.

Again, we need to reduce the number of attributes to 7. The results of the experiment are shown in Fig. 20.

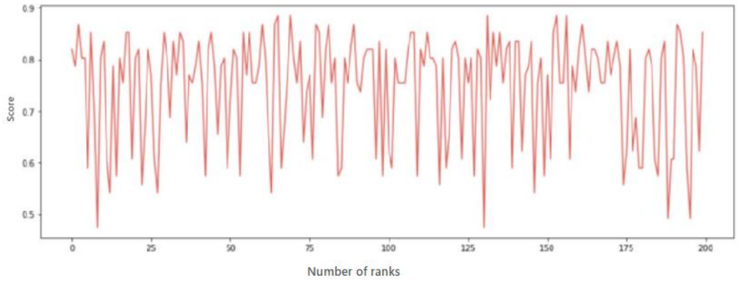


**Figure 20**: Average classification scores using 7 attributes

Fig. 20 x-axis indicates the number of random states, y-axis - the average score for this method. The figure shows that the score ranges from 0.5 to 1.

Thus, the average estimate of the Naive Bayes classifier for random states from 0 to 200 is 0.765245901639342.

## 4.3. Application of the Multinomial Naive Bayes classifier

To implement the classification, you should use MultinomialNB from the sklearn library with different states when sharing data.

The score function from the sklearn library is used to evaluate the results.

The obtained results are presented in Fig. 21.

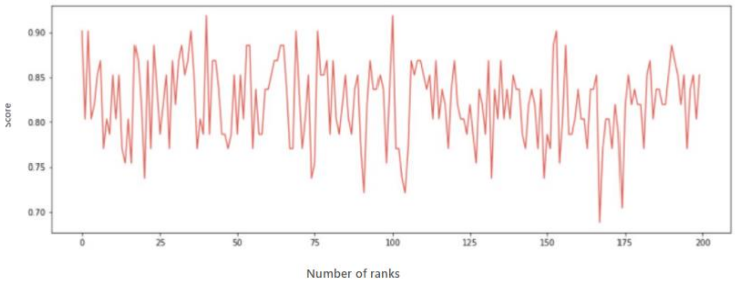


**Figure 21**: Average classification scores using all attributes

In Fig. 21 x-axis indicates the number of random states, y-axis - the average score for this method. The figure shows that the score ranges from 0.7 to 1.

Thus, the average score of the Multinomial Naive Bayes classifier for random states from 0 to 200 is 0.850129016334426.

To illustrate the algorithm's performance, you need to build a matrix of inconsistencies (confusion matrix) (Fig. 22).

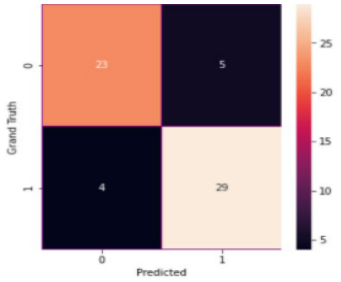
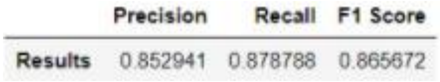


**Figure 22**: Matrix of discrepancies

It is also necessary to determine the accuracy of Fig. 23.



|  | Precision | Recall | F1 Score |
|---|---|---|---|
| Results | 0.852941 | 0.878788 | 0.865672 |

**Figure 23**: Measures of accuracy

Fig. 23 shows the accuracy or positive predictive value (precision), probability of detection (recall), and completeness of determination (f1_score).

The next step is to reduce the number of attributes to 10. You need to remove the parameters that have the most negligible impact on heart disease and apply the Multinomial Naive Bayes classifier again. The results of the experiment are shown in Fig. 24.

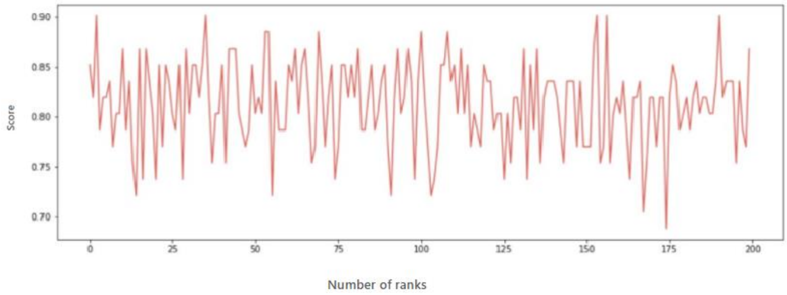


**Figure 24**: Average classification scores using 10 attributes

In Fig. 24 x-axis indicates the number of random states, y-axis - the average score for this method. The figure shows that the score ranges from 0.7 to 1.

Thus, the average estimate of the Multinomial Naive Bayes classifier for random states from 0 to 200 is 0.82. The highest score is 0.849039016334426.

The next step is to reduce the number of attributes to 7. The results of the experiments are shown in Fig. 25.

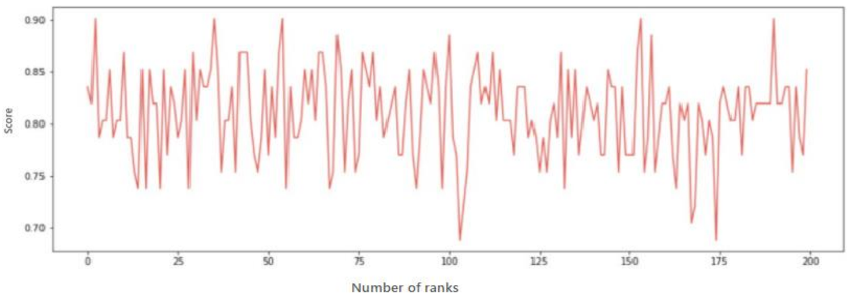


**Figure 25**: Average classification scores using 7 attributes

Fig. 25 x-axis indicates the number of random states, y-axis - the average score for this method. The figure shows that the score ranges from 0.7 to 1.

Therefore, the average estimate of the Multinomial Naive Bayes classifier for random states from 0 to 200 is 0.83010.

## 5. Discussion of experimental results

To study the accuracy of the two classification models, we use a set of data on heart disease.

Table 1 summarizes the characteristics of the data set used in the experiments.

**Table 1**
Data set characteristics

| Dataset | Examples | Train data | Class | No. of features |
|---|---|---|---|---|
| Heart | 303 | 240 | 2 | 14 |

Table. 1 shows the use of the Heart data set, which contains 303 data sets, of which 240 are used for training. The data set includes two classes and 14 characteristics.

Table. 1 mentions the features that have been tested to training algorithms.

The data set was initially studied using 14 features. Subsequently, ten features were selected and rejected those that had the least impact on heart disease. And finally, seven features. The training of the three classification models are given in Table. 2.

**Table 2**
Dividing data to test and training

| Method | Dataset (Heart) | |
|---|---|---|
| | Train | Test |
| GaussianNB | 80% | 20% |
| MultinomialNB | 80% | 20% |

Table. 2 shows the results of data sharing for training and testing. Both algorithms obtained data that were equally separated.

We chose an 80:20 ratio because the Naive Bayes classifier could not benefit from retraining the data.

**Table 3**
Comparison of accuracy of two classifier models

| Method | Accuracy % | | |
|---|---|---|---|
| | 14 features | 10 features | 7 features |
| GaussianNB | 0,84426229 | 0,83032786 | 0,765245901 |
| MultinomialNB | 0.85012901 | 0.849039016 | 0.830100260 |

Table. 3 shows the results of classification, the accuracy of different models of Naive Bayes. In the study, Multinomial Naive Bayes achieved the highest average accuracy with 0.85%. This shows that the multinomial classifier surpassed the Gaussian model. Fig. 26 shows a comparison of the accuracy of the two methods.

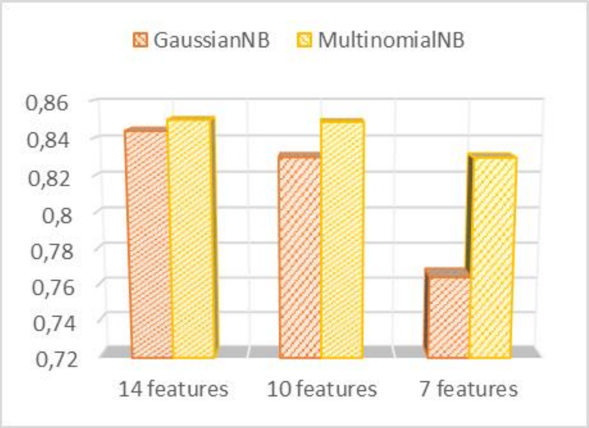


**Figure 26**: Comparison of the accuracy of two methods of the Naive Bayes classifier

In Fig. 26, orange depicts a Naive Bayes classifier with a Gaussian distribution. The Naive Multinomial Bayes classifier is shown in yellow. The x-axis indicates the number of features used for the experiments. The y-axis shows the achieved accuracy.

From Fig. 26, it is noticeable that both methods show worse results when the number of features decreases. This is because we first rejected the symptoms, which had little effect on heart disease. Therefore, the accuracy is similar.

In addition, we can notice that the Multinomial classifier shows much better results when reducing the number of features. This advantage is because this method makes the second assumption of positional independence. Conditional probabilities of symptom onset are equally independent of its position in the data sample.

Let's take into account the nature of the chosen problem in the study, namely the values of "0" and "1" in the answers of the classifiers. We can conclude that the correct classification of first-class objects is, in our case, more critical. After all, it is better to do all the tests once again for a healthy person than not to recognize the disease in a sick person.

That is why it is worth emphasizing the recall score when comparing models with each other and choosing the best one. It estimates the proportion of correctly classified first-class objects. In addition, it is necessary to reach the positive predictive value (precision) and the completeness of the definition (f1_score) (Table 4).

**Table 4**

Comparison of evaluations of two classifier models

| Method | Accuracy % | | |
|---|---|---|---|
| | Recall | Precision | F1_score |
| GaussianNB | 0,868788 | 0,828571 | 0,852941 |
| MultinomialNB | 0,878788 | 0,852941 | 0.865672 |

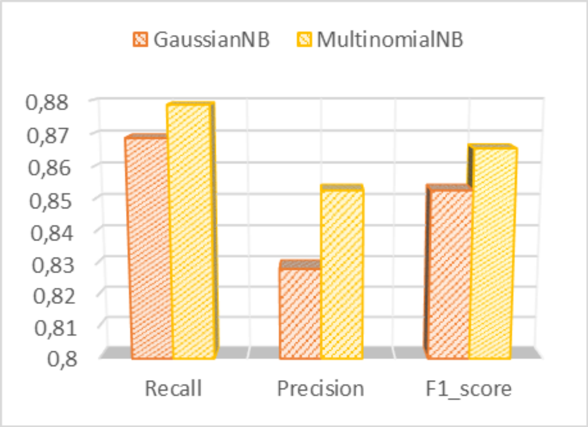


**Figure 27**: Comparison of estimates of two methods of the Naive Bayes classifier

In Fig. 27, orange depicts a Naive Bayes classifier with a Gaussian distribution. The Multinomial classifier is shown in yellow. The x-axis indicates the selected estimates used for the experiments. The y-axis shows the achieved accuracy.

Analyzing the figure, we can conclude that with the help of the Multinomial Bayes classifier, the number of sick patients in whom the disease will be detected is more significant. Using this classification method, more people will receive a correct diagnosis and, therefore, will have a chance for treatment and recovery.

We also compare the operating time of the two classification methods. Namely, determine the time of training (Table 5).

**Table 5**
Division of data into test and training

| Method | Time, s | | |
|---|---|---|---|
| | 14 features | 10 features | 7 features |
| GaussianNB | 0,01301 | 0,00902 | 0,00603 |
| MultinomialNB | 0,01196 | 0,00881 | 0.00399 |

Table 5 shows the execution time of the classification of different Naive Bayes models. On the same data set, MultinomialNB performs training faster, which again emphasizes its advantage for the selected data set.

It is also noticeable that as the number of features decreases, the time decreases (Fig. 28).

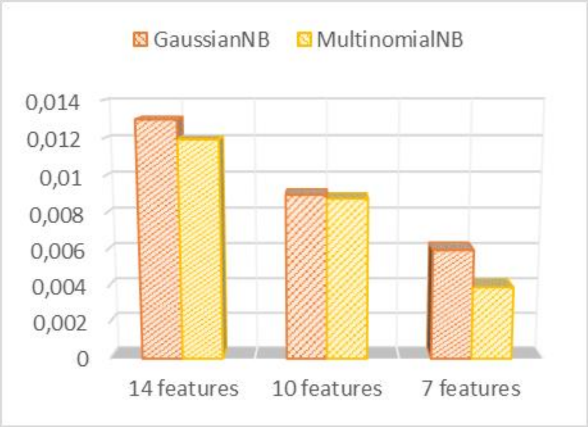


**Figure 28**: Comparison of learning time of two methods of the Naive Bayes classifier

In Fig. 28, orange depicts a Naive Bayes classifier with a Gaussian distribution. The Multinomial classifier is shown in yellow. The x-axis indicates the number of features that were used for the experiments. The y-axis indicates the training time.

Analyzing Fig. 28, we can conclude that the Multinomial Bayes classifier is more accurate and faster for the selected data set.

So, the choice of using the Naive Bayes method depends on the data. The Multinomial Naive Bayes is appropriate if the data consists of calculations, and observations can only take non-negative integers. It is better to use the Gaussian NB for decimal features. GNB accepts features that correspond to the normal distribution.

For the selected data set, which contains features for diagnosing heart disease, the Multinomial Naive Bayes showed better results. Using this method, we can achieve greater accuracy and reduce the time to perform training.

Analyzing the study results, it is worth emphasizing the importance of choosing the correct method of the naive classifier. It helps achieve better classification results, which is critical in the medical field.

## 6. Conclusion

The paper considered the relevance of the topic: the use of data mining methods for diagnosing the disease in a patient on a set of indicators, such as symptoms, test results, and other indicators.

We used the Heart data set for the study, which we cleared of emissions, Null values, and normalized. We also performed a search and analysis of significant features and patterns between different factors influencing heart disease.

In addition, we used two algorithms in this work, which objectively showed the classification results on the selected dataset.

The parameters used for the analysis were the selection and deletion of the function. We first tested a classifier with all the features and then gradually reduced the set to determine which algorithm best classifies with fewer features.

The simulation results show that the Multinomial Naive Bayes classifier has better accuracy than the Gaussian method with the same data set and parameters. In addition, it reduces training time, which is very important because the annual growth of data in medicine is increasing very rapidly.

In future work, it is worth considering two aspects. Namely, we can compare more algorithms to achieve better results and potentially introduce a better algorithm in Naive Bayes. Moreover, we can try to evaluate the effectiveness of their work to justify their use in the health care system.

## 7. References

[1] G. Parthiban, A. Rajesh, S.K.Srivatsa, Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method, in: International Journal of Computer Applications, 24(3) (2011). doi: 10.5120/2933-3887.

[2] G. Subbalakshmi, K. Ramesh, M.Tech, M. Chinna Rao, M.Tech, Decision Support in Heart Disease Prediction System using Naive Bayes, in: Indian Journal of Computer Science and Engineering, 2(2) (2011):170-176.

[3] J. Soni, U. Ansari, D. Sharma, S. Soni, Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, in: International Journal of Computer Applications, Vol. 17(8) (2011): 43-48. DOI:10.5120/2237-2860.

[4] Ch. Vincy, M.S. Bindu, Prediction Analysis of Cardiac Disease using Classification (2019). doi: 10.22214/ijraset.2019.6295.

[5] N. Kunanets, O. Vasiuta, N. Boiko, "Advanced Technologies of Big Data Research in Distributed Information Systems", in: Proceedings of the 14th International conference "Computer sciences and Information technologies" (CSIT 2019), Lviv, Ukraine, September 17-20 (2019): 71-76. DOI: 10.1109/STC-CSIT.2019.8929756.

[6] Heart Database. URL: https://www.kaggle.com/zhaoyingzhu/heartcsv.

[7] Clinic Manufactory - Cardiovascular Diseases. URL: https://manufacturaclinica.com/blog/sertsevo-sudinni-zahvoryuvannya.

[8] W.J. Loesche, Periodontal disease as a risk factor for heart disease, in: Compendium, 15(8): 978.

[9] S. Kharya, S. Soni, Weighted naive bayes classifier: A predictive model for breast cancer detection, in: International Journal of Computer Applications, 133(9) (2016): 32-37.

[10] A. Ashari, P. Iman, and A. Min Tjoa, "Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool", in: International Journal of Advanced Computer Science and Applications (IJACSA) (2013).

[11] N.D. Uma, Extraction of action rules for chronic kidney diseas using Naive Bayes classifier, IEEE Internstional Conference Comput Intelligence Comput Res (2016).

[12] W. P. Castelli, Lipids, risk factors and ischaemic heart disease, Atherosclerosis (1996). doi: 10.1016/0021-9150(96)05851-0.

[13] W. F. Wilson, W. B. Kannel, H. Silbershatz, Clustering of Metabolic Factors and Heart Disease. 159(10) (1999): 1104. doi: 10.1001/archinte.159.10.1104.

[14] Stat Quest with Josh Starmer - Naive Bayes. URL: https://www.youtube.com/watch?v=O2L2Uv9pdDA&ab_channel=StatQuestwithJoshStarmerStatQuestwithJoshStarmer%D0%9F%D1%96%D0%B4%D1%82%D0%B2%D0%B5%D1%80%D0%B4%D0%B6%D0%B5%D0%BD%D0%BE.

[15] N. Boyko, Kh. Shakhovska, L. Mochurad, J. Campos, “Information System of Catering Selection by Using Clustering Analysis”, in: Proceedings of the 1st International Workshop on Digital Content & Smart Multimedia (DCSMart 2019), Lviv, Ukraine, December 23-25, (2019): 94-106.