ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)*

UDC 004.8 DOI: https://doi.org/10.30837/2522-9818.2024.28.076

N. HULIIEV

CHOICE OF MACHINE LEARNING MODELS FOR PREDICTING THE DEVELOPMENT OF PSYCHOLOGICAL DISORDERS IN PEOPLE WITH HYPOTHIREOSIS AND HYPERTHIREOSIS

The **subject** of this article is endocrinological diseases, namely, the analysis of complications in people with hypothyroidism and hyperthyroidism. It is known that these diseases occur asymptomatically or in a way that may indicate other possible diseases, so people do not suspect what exactly they are suffering from. Later, the diseases develop to the point where complications occur in the body, some of the most dangerous of which are psychological disorders: depression, mania, aggression, etc. Therefore, the **aim** of this work is to develop methods for predicting the occurrence of neurological deterioration in people who have already been diagnosed with endocrinological diseases. The article solves the **problem** of choosing the best models for predicting the occurrence of psychological disorders in people with endocrinological problems. Machine learning methods that are widespread in the medical field were analyzed and one of them was chosen that more optimally solves all the tasks of the task. The selection of criteria took into account potential problems with medical and psychological data. The **method** used was linear additive convolution, which is used to select the best alternatives according to the results, with the Pareto principle, which aims to exclude less suitable alternatives because all the features have lower values than in other options. For the experiment, all features were converted into quantitative ones to calculate convolution values. The evaluation criteria are given in the paper. The following **results** were obtained: the forecasting model in further study of this problem will be a random forest. **Conclusions**: the forecasting methods were studied and a more optimal model was chosen using linear additive convolution, namely, the random forest algorithm, its advantages and disadvantages were considered. A more detailed analysis of its development will be presented in the following articles. A mathematical description of the chosen forecasting method is provided, which includes potential ways of implementation and steps for building an algorithm for one of these methods.

Keywords: hypothyroidism; hyperthyroidism; psychological disorders; forecasting; linear additive convolution; Pareto principle; random forest; decision tree; Gini index.

Introduction

Hypothyroidism and hyperthyroidism are among the most common endocrine diseases of the thyroid gland, the factors of which are

– environment
– bad habits;
– genetics;
– unhealthy diet;
– allergies;
– iodine deficiency;
– stress.

According to the World Health Organization, they rank second among endocrinopathies, and diabetes mellitus ranks first.

According to research, the total number of people who develop manifest hypothyroidism ranges from 3% to 8%, and if we add cases of subclinical hypothyroidism, it is 10% to 12%. The consequences of this disease increase over time, as it has many effects on various organs of the patient. Most often, the cardiovascular and nervous systems are affected. The disease negatively affects physical, sexual, cognitive, and intellectual functioning, and therefore may have atypical symptoms that force patients to see different doctors because of concerns about the heart, nerves, stomach, and reproductive function, without realizing that the problem is different, and local treatment will not help overcome a global disease.

A significant number of observations of endocrine disorders are devoted to the study of patients' psychological health, as it is known that deterioration can range from passive or increased exhaustion to unexpected aggressive and dangerous actions. And a long course of the disease can cause parts of a person's personality to be removed, such as states of affect and memory impairment.

All patients develop complications that lead to a deterioration in the psycho-emotional state, the nature of which may vary depending on the severity of the disease. M. Bleuler studied mental syndromes of a narrow circle, namely deviant behavior that occurs

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

due to endocrine diseases. His systematization is that the scientist combined them into one structure, which includes the following components:

– mood deterioration;
– lack of motivation;
– decrease in activity;
– change in instincts;
– change in urges [1].

Hyperthyroidism is a disease in which the thyroid gland produces excessive amounts of hormones, which can cause thyrotoxicosis. The disease occurs due to the following causes: toxic denoma, toxic multinodular goiter, and Graves' syndrome. Diagnosis of the disease involves the use of imaging methods, ultrasound, monitoring of iodine absorption, and biochemical tests. Thyroid dysfunction negatively affects the skin, reproductive, cardiovascular, immune, and gastric systems. Treatment options include antithyroid drugs, surgery, and radioactive iodine.

The disease can occur suddenly or develop in the body over time, go away on its own or in remission. The following symptoms are subtle: ventricular arrhythmias, tachycardia. Dangerous signs are disorders of the skin, eyes, musculoskeletal system, and neck. Other signs of the disease include insomnia, anxiety, irritability and depression, memory and attention impairment, delirium, and apathy. One of the less common complications is psychosis.

The relevance of the analysis of these two diseases is that their rapid spread among the population, asymptomatic or ambiguous development (i.e., the very signs that may indicate abnormal endocrine behavior of the body) make us believe that other problems have arisen. After all, due to the many forms they can take, it is very difficult to diagnose them in time, which causes further complications, namely psychological disorders. The patient's psycho-emotional state deteriorates, ranging from various syndromes to severe disorders. Therefore, after the diagnosis of hypothyroidism and hyperthyroidism in a patient, it is necessary to immediately analyze his or her mental health in order to prevent its noticeable deterioration or the occurrence of diseases associated with the nervous and mental systems of the body in a timely manner.

Analysis of recent research and publications

Machine learning is becoming a widespread means of prediction and diagnosis in medicine. There are many methods aimed at determining the probability of certain signs occurring given certain circumstances, which is the most common purpose of their use in this area. Therefore, the study of diseases using data mining methods is currently appropriate and relevant [2, 3].

When discussing diseases that are common in the world, endocrinological diseases are still the ones that affect humanity. It is known that hypothyroidism and hyperthyroidism, which are thyroid disorders that cause abnormal regulation of thyroid hormones, can develop asymptomatically or with signs that indicate completely different diseases, becoming factors in other complications due to late diagnosis [4, 5].

Some of these deteriorations are psychological disorders, such as depression, mania, aggression, etc. Neurological problems increase the damage to the entire body, which makes it impossible to treat a patient with hypothyroidism or hyperthyroidism normally [6]. Currently, many observations are aimed at analyzing the course of endocrinological diseases or at options for mitigating symptoms, so the problem of preventing the development of psychological disorders due to hypothyroidism and hyperthyroidism remains relevant. This paper focuses on selecting one of the most optimal machine learning methods for predicting whether neurological problems may occur due to endocrinological diseases [7].

Currently, a large number of models are used in the medical field to improve diagnostic and preventive treatment. Unfortunately, there are still cases when it is extremely difficult to determine what exactly is affected in the body, what are the causes of the disease, neoplasms, complications, and even the patient's condition. In these difficult tasks that doctors around the world have to solve, it has become advisable to use prognostic models in order to act in a timely manner to find the appropriate and correct treatment for a person as soon as possible, because it is easier to solve problems before they occur.

One of the cases of building predictive models is the development of a mathematical model to predict intrauterine infection among newborns, as the obstetrics and gynecology department warns of an increase in the number of intrauterine infections that disrupt pregnancy and increase the likelihood of prenatal death. A projective and retrospective study was conducted among women with viral, bacterial, and combined infections. The analysis included clinical and obstetric examinations, and the prediction was made on a two-point scale using multivariate discriminant analysis using the *Statistica* statistical environment based on 105 indicators that were

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

the result of a medical examination. As a result, the most likely and influential factors were identified by the factor structure matrix of the discriminant analysis protocol.

Pregnancy is known for many unpredictable problems, so this topic is common among scientists. Another study was aimed at creating a model for predicting the course of pregnancy depending on women's laboratory and instrumental parameters to reduce the likelihood of preterm birth or antenatal fetal death. Initially, all women underwent obstetric and gynecological and somatic examinations to collect materials for analysis. Means and standard deviations were calculated using standard methods, but if there was a significant discrepancy between the values, median and quartile values were also calculated. Pearson and Mann–Whitney's tests of agreement were used to calculate the reliability of sample differences. Full statistical analysis was performed using the *Statistica* 6.0 software package. The predictive model was developed using fuzzy logic based on the Takagi-Sugeno fuzzy framework. As a result, four models were built to predict the term of labor, which is not likely to be preterm and possible threats that, on the contrary, can cause it.

Hypothyroidism and hyperthyroidism are also being studied using machine learning. For example, representatives of Kharkiv National Medical University analyzed the course of primary hypothyroidism in Ukrainians who were forced to leave their homes due to the war, which became a factor in cognitive and anxiety disorders. Changes in the life of internally displaced persons are accompanied by a modification of social and psychological relations, which negatively affects the mental state of a person, exacerbating the development of depression. IDPs and persons permanently residing in the Kharkiv region with a diagnosis of primary hypothyroidism participated in the study. After clinical-neurological and clinical-psychopathological analyzes, the results were processed using a mathematical and statistical approach using *Statistica* 6.0 and Student's *t*-test. As a result, the average values with a possible arithmetic mean error were obtained, and the dependencies between the values were determined using correlation analysis. It was found that among IDPs there are more people with primary hypothyroidism manifested in depression of varying degrees than in the second group of people. Also, thyroid hormone deficiency caused anxiety in all, but in those who did not change their place of residence during the war, it was found to be the most common [8–10].

It is worth mentioning studies on the treatment of hypothyroidism. The Ivano-Frankivsk Medical University considered the question of the greatest effectiveness of alpha-lipoic acid Dialipon or the drug Vitaxon. The medical parameters of 42 middle-aged patients with primary hypothyroidism were studied: clinical signs, neuropsychiatric disorders, organ damage. Statistical processing was performed using the *Statistica* package (*StatSoft, Inc*.) and nonparametric methods of evaluating the results. Patients were divided into two groups: the first group included people who were to take Dialipone and Vitaxon as hormone therapy, and the second group was prescribed *L*-thyroxine-randomly. Some patients from both groups got better, but a noticeable improvement was observed among people taking Dialipon and Vitaxon. In addition, the liver also showed positive dynamics, because it is responsible for controlling the body's metabolism. Therefore, given that metabolic changes can provoke nervous system damage, it is necessary to take additional medications [11, 12].

An interesting study of endocrinological disease was conducted on 48 mice. During the observation, the effect of stress and physical activity on the thyroid gland in the setting of hypothyroidism was analyzed. We directly studied how chronic stress and physical training can change the morphology of the gland by means of microscopy and statistical analysis based on the Student's *t*-test method. As a result, it was found that the effects of stress and exercise did not change the number of iodine-containing hormones and thyroid TSH in hypothyroidism.

Endocrinopathies also have an impact on the dental health of patients. Studies show that hypothyroidism and hyperthyroidism cause pathological processes in the periodontium, caries, and non-carious formations. These processes are caused by the fact that thyroid diseases disrupt metabolism, which provokes enamel and dentin erosion, enamel necrosis, and tooth abrasion. In patients with thyroid dysfunction, there is a correlation between the prevalence of periodontal disease (generalized periodontitis) and the time of disease development and the activity of the process [13–15].

Identification of previously unsolved parts of the overall problem. Purpose of the work, tasks

Currently, observations of known endocrinopathies are devoted to the study of the course, possible treatments and features of further complications, but the problem

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28) ISSN 2524-2296 (online)

of preventing one of the most common complications – the development of psychological disorders among people with hypothyroidism and hyperthyroidism – still remains relevant. The purpose of this study is to select the most optimal or optimal methods for predicting the deterioration of patients' psychological health, which will process (process) medical and psychological indicators of patients. And the next step will be to analyze the proposed ways to avoid these difficulties against the background of endocrinological disease [7].

Methods and materials

The multi-criteria task of choosing the most optimal forecasting method in the medical field is to determine the best one among all the proposed ones. Two types of methods are used for this purpose. The first set of methods aims to remove the number of evaluation criteria by making assumptions in the process of ranking the values of characteristics, and the second removes possible options before the comparison begins.

The most effective method for this observation is still the method from the first set, which includes the convolution method, boundary criteria, distance, and the main criterion.

The convolution method summarizes all the criteria. Such methods are divided into additive, multiplicative, and maximin convolution.

Additive is calculated by the formula

$$K(x)=\sum_{j=1}^{n} a_j K_j(x), \qquad (1)$$

where $K(x)$ –general criterion for the alternative $x \in X$; $\left(K_1(x),\ldots K_j(x),\ldots K_n(x)\right)$ – a set of initial criteria; n – number of initial criteria; a_j – a normalization factor indicating the weight of the alternative.

The best of all possible alternatives to the problem is calculated using the following formula:

$$x^* = \arg\max_{x\in x} K(x) \ . \qquad (2)$$

That is, the result is the largest value obtained by the convolution method.

Multiplicative convolution is calculated using the formula

$$K(x) = \Pi_{j=1}^{n} K_j^{a_j}(x). \qquad (3)$$

The maximin convolution is calculated by the formula

$$K(x) = \max_i \min_j a_{ij} K_j(x). \qquad (4)$$

The best results for the multiplicative and maximin convolutions are calculated using formula (2).

The method of threshold criteria is used in design and planning problems in which the threshold values of the criteria take on the values $k_j(x) \ge k_{j^0}$; $j = 1\ldots, n$. The calculation formula for this method is as follows:

$$K(x) = \min_j \left(K_j(x)/K_{j0}(x)\right). \qquad (5)$$

The best result is selected by formula (2).

The distance method uses distance as an additional metric. For example, the following information is enough to select the ideal solution $(K_0, ..., K_{0n})$. Let's calculate the distance to the maximum value $d(x)$ for each alternative. Then the best alternative will be determined by the formula

$$x^* = \arg\min_{x\in x} d(x). \qquad (6).$$

In working with the methods of the first group, methods from the second group are used, namely, the Pareto principle, when the best option is selected from the list of alternatives remaining after this method has eliminated the others by comparing their characteristics and identifying the worst options because their values had lower indicators.

The principle of equilibrium, or Nash's principle, aims to reduce the number of alternatives and calculates which one is inferior in terms of characteristics to the others; it is closely related to the Pareto principle.

However, there are cases when uncontrollable parameters complicate the solution of multi-criteria problems, which can arise for various reasons. In such cases, it is advisable to use the guaranteed outcome method, which allows you to determine the worst case response and a likely high and guaranteed value.

By considering the methods from the two groups, you can choose the method that is most effective for the study. It is not advisable not to analyze all the options, so we will use the method from the first group and convolution, because it is difficult to determine the thresholds of the criteria.

The most common and simplest option is linear additive convolution, so we will use it as a method for determining the usefulness of models to select the best one.

The first step is to identify all the criteria that will be involved in the analysis of alternatives, and then calculate their weighting values to describe the priority in choosing the best option.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

The values of each criterion can be quantitative or qualitative. This method operates with the former, so if the alternatives have values of the latter type, it is necessary to convert them to quantitative values.

When the quantitative values are known, alternatives can be eliminated from the list using the Pareto principle if the values of all criteria for a particular alternative are lower than those of other options. Next, the indicators are normalized if they are in different ranges or measures of measurement, which can lead to inaccurate and incorrect results, so it is better to normalize the values of all criteria of all alternatives with a range from 0 to 1.

In the case of maximization, you can normalize by dividing the value of the criterion by the maximum, while in the case of minimization, one is divided by it.

The next step is to rank the criteria for calculating the weighting coefficients. We have n criteria, of which the best one will have a value of n divided by n, the least important one will have a value of $n-1$ divided by n, and so on. Another way is to divide one by the sum of all the criteria scores.

The last step is to calculate the convolution value based on the alternatives: calculate the sum of the products of the criteria values and their weighting factors [16].

We have investigated forecasting methods for selecting the best model for observation purposes. To predict the development of mental disorders in patients with hypothyroidism and hyperthyroidism, it is necessary to apply methods that take into account the medical and psychological indicators of patients. Further research is aimed at analyzing the requirements for this task and selecting the most optimal model among all those described in the table.

The task is to solve a multi-criteria problem, namely, to determine which machine learning method will better predict the possible development of psychological disorders in people with hypothyroidism and hyperthyroidism to identify future ways to prevent them.

First, let's define a set of alternatives – these are models that are more commonly used in the medical field, among which we will choose an effective one for the study.

Let's assume that we have

– linear regression;
– polynomial regression;
– logistic regression;
– decision trees;
– multilayer perceptron;
– *k-nearest* neighbors model;
– *random forest*;
– gradient boosting;
– Bayesian classification;
– ensemble of models;
– SVM.

Medical and psychological indicators of a patient are determined by their unstable nature. They can have abnormal, incomplete, empty, nonlinear values in a rather significant amount, because psychological indicators will be accurately taken by questionnaires, interviews, and non-verbal tests. Therefore, the best option should not neglect the peculiarities of these indicators. This means that the model must process a large amount of input information, which may be nonlinear, given that the model must be able to handle missing indicators and respond to noise. Therefore, the selection criteria were as follows:

– model complexity;
– type of training;
– ability to process nonlinear information;
– whether the error is taken into account;
– tendency to overlearn;
– working with large amounts of information;
– working with missing information;
– work with noise.

Create and fill in a table with all the alternatives and the criteria that describe them (see Table 1).

Next, you need to convert the value of the criteria into a numerical value. Let's consider each of them.

The complexity of the model lies in the training method, the complexity and number of algorithms in one forecasting method, and the number of layers in the case of neural networks. Therefore, the values "simple", "medium", and "complex" are given accordingly.

The type of training presented in the table means "with a teacher" or "without a teacher". It is easier to create a model that does not require time to learn and search for information, but if learning is based on previous information, the result may be more likely, so if the value is "without a teacher", it is 1 point, and if it is the other way around, it is 2 points.

The criterion "ability to process nonlinear information" indicates whether the model is able to work with indicators that are scattered nonlinearly, as they may have unexpected values, which is likely to affect the overall result. Therefore, this feature should be taken into account in the study. If the model has this feature, the value is 1 point, if not, it is 0.

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Table 1. *Experimental results*

Types	Characteristics							
	Simplicity of the model	**Type of training**	**Ability to process nonlinear information**	**Does they take into account an error?**	**Tendency to relearn**	**Working with large amounts of information**	**Working with missed information**	**Working with noise**
Linear regression	Simple	With a teacher	No	No	Less tend	Yes	Can not	Can not
Polynomial regression	Simple	With a teacher	Has	No	Tend to relearn	Works, but may be problems	Can not	Can not
Logistic regression	Simple	With a teacher	No	No	Less tend	Yes	Can not	Can not
Decision trees	Medium	With a teacher	Has	No	Tend to relearn	Works, but may be problems	Can	Can
Multilayer perceptron	Medium	With a teacher	Has	No	Tend to relearn	Yes	Can not	Can not
Model *k-nearest* neighbors	Simple	Without a teacher	Has	No	Tend to relearn	Small amount of information	Can not	Can not
Random Forest	Medium	With a teacher	Has	Yes	Tend to relearn	Yes	Can	Can
Gradient boosting	Complex	With a teacher	Has	Yes	Less tend	Works	Can not	Can
Bayesian classification	Simple	Without a teacher	Has	Yes	Less tend	Works	Can not	Can not
Ensemble of models	Complex	With a teacher	Has	Yes	Tend to relearn	Works	Can	Can
SVM	Medium	With a teacher	Has	No	Less tend	Works, but with limitations	Can not	Can not

Usually, a model cannot predict the exact percentage of probability, so it is advisable to use a method that takes into account the measurement error. If the model does, the value is 1 point, otherwise it is 0 points.

Each model may be prone to overfitting, which is possible under different conditions, or not at all, so if the method is likely to have this problem, which is the worst, the parameter value is 0, if it is likely to have it under special conditions, it is 1, and if the probability of this is low (which is the best), then it is 2 points.

One of the main characteristics is working with large amounts of information, so if this is natural for the model, it is 2 points, if there are some restrictions, then it is 1 point, otherwise - 0.

If the model can work with missing information or noise, then it gets 1 point, and otherwise 0 points.

Replace the values in the table with quantitative indicators. Also, at this stage, some alternatives can be eliminated according to the Pareto principle if they are inferior to other options by the criteria (see Table 2).

Table 2. *Modified table after changing the information with quantitative indicators and applying the Pareto principle*

Types	Characteristics							
	Simplicity of the model	**Type of training**	**Ability to process nonlinear information**	**Does they take into account an error?**	**Tendency to relearn**	**Working with large amounts of information**	**Working with missed information**	**Working with noise**
Linear regression	3	2	0	0	2	2	0	0
Logistic regression	3	2	0	0	2	2	0	0
Random Forest	2	2	1	1	1	2	1	1
Gradient boosting	1	2	1	1	2	2	0	1
Bayesian classification	3	1	1	1	2	2	0	0
SVM	2	2	1	0	2	1	0	0

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)

The last step is actually to calculate the values of the linear additive convolution for each option, with the value of the normalization factor calculated first for each criterion (see Table 3).

Table 3. *Convolution results*

Model	Convolution value
Linear regression	0,75974026
Logistic regression	0,75974026
Random Forest	2,680735931
Gradient boosting	1,70021645
Bayesian classification	1,252164502

As we can see, according to the convolution results, the best model for this study is the random forest algorithm.

Research results and discussion

According to the results of the study, the best machine learning method to be used to predict the development of psychological disorders among people with hypothyroidism and hyperthyroidism is *Random Forest* [16].

Decision trees are known for their overfitting, which causes an increase in the variance of predictions. The *Random Forest* algorithm was developed to solve the above problem, allowing to build ensemble forecasts, but with a lower variance value, and it is similar to backpropagation. Random forest is a modified decision tree algorithm aimed at building not one but many trees, each of which produces a certain result, and the final one is the one that occurs most often. However, it differs in that it has a second level of randomness: in the process of optimizing node crushing, a random subset of features is analyzed for subsequent decoration of the estimators, and the random forest always determines the size of the bootstrapped data set, according to the size of the training sample [17].

Random forest significantly increases the accuracy and efficiency of forecasting and classification. The algorithm works as follows: first, during training, a tree based on random information is built, and in the process of dividing the nodes, a random subset of characteristics is selected and the result that occurs most often becomes the final one.

The advantages are:

– nonlinear information does not affect the efficiency of the algorithm;

– support for parallel processing;

– simplicity of application is that the only parameters of the method are the number of randomly selected features and the number of trees to be built on a randomly selected subset of the data sample;

– there is no need to reduce the tree;

– the algorithm estimates the importance of criteria and out-of-band accuracy in programs with large amounts of information, where estimates can be overestimated.

But like bagging, random forest is not defined by a lower bias. If a significant amount of information contains unequally distributed and mutually independent examples, overfitting is used – the process of selecting many identical decision trees by the random forest algorithm, each of which is overfitted, which is a known drawback.

In addition, it is prone to overfitting under certain conditions, namely, if there are too many trees in the forest, high correlation between them, small sample size, incorrect set of hyperparameters, and too complex data. The following methods are used to reduce the likelihood of overfitting:

– cross-validation;

– increasing the size of information;

– limiting the depth of trees;

– tree diversity.

Despite its drawbacks, the *Random Forest* algorithm is a more optimal option that processes a significant amount of information, is able to work with noise, and processes nonlinear, missing data, including measurement error. It has several implementation methods [18, 19].

The algorithm is used to evaluate the importance of the characteristics that need to be trained based on the average *out-of-bag* error for each subsample item. Next, before and after shuffling, it is necessary to determine the average value of the difference in *out-of-bag* errors on all trees, normalized by the standard deviation.

The main thing in building decision trees is the method of selecting the attribute by which the division will take place and the nodes will be built. There are the following methods:

– ID3 algorithm, which uses the Gini index or incremental method;

– C4.5 algorithm, which is a better version of ID3, which takes into account the normalized growth;

– CART algorithm;

– modifications of the CART algorithm – IndCART, DB-CART.

All trees are built in the following independent steps:

– generate a subset of size n from the training data set randomly;

– build a tree of m randomly selected features;

– continue the process without cutting off until the amount of data is complete.

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

In general, the algorithm for constructing a decision tree is as follows: it is necessary to calculate the entropy of the input set s_0, if $s_0 = 0$, then:

1) all sample objects are of the same class;

2) store this class as a leaf of the tree.

If $s_0 != 0$, then:

1) determine the attribute that will divide the set in such a way as to reduce the average entropy value;

2) the found attribute becomes a node of the decision tree and is saved;

3) divide the sample into subsets depending on the values of the selected attribute;

4) recursively continue the process for each subset [20, 21].

Let's consider one of the tree building algorithms – the CART algorithm. In it, each node has two subnodes. At each step, the selected node attribute divides the set into two parts: the right part, in which the rule is executed, and the left part, in which the rule is not executed. To select the optimal rule, the partitioning quality evaluation function is applied. The evaluation function, which uses the CART algorithm, is based on the intuitive idea of reducing uncertainty in a node. This means a partitioning that will result in a node having as many examples of one class as possible and as few as possible of all other classes. This concept is close to entropy, but it uses a different measure of uncertainty, for which the term "dirty node" is appropriate. In the CART algorithm, the idea of a "dirty node" is formalized in the *Gini* index. If a data set T contains data from n classes, then the *Gini* index is defined as

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2 , \qquad (7)$$

where the parameter p_i – is the probability of class i in T.

If the set T is split into two parts, T_1 and T_2 with the number of examples in each N_1 and N_2 respectively, then the quality index of the split is equal to

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2). \qquad (8).$$

The best split is the one for which $Gini_{split}(T)$ is minimal. We denote the number of examples in a node as N, where L and R are the number of examples in the left and right descendants, respectively, l_i and r_i are the number of examples of the i-th class in the left/right descendant. Then the quality of the partitioning is estimated by the following formula:

$$Gini_{split}(T) = \frac{L}{N}\left(1 - \sum_{i=1}^{n}\left(\frac{l_i}{L}\right)^2\right) + \frac{R}{N}\left(1 - \sum_{i=1}^{n}\left(\frac{r_i}{R}\right)^2\right) \to \min. \quad (9)$$

The peculiarity of this index is the most optimal breakdown of data to build a better decision tree [22, 23].

Conclusions and prospects for further development

Modern medicine is increasingly using machine learning methods to diagnose and predict diseases, their course, and types of treatment. Such methods are becoming widespread in this field, as they increase the chances of a safer, more desirable, and accurate outcome. Specifically, endocrinopathies are known for being extremely difficult to detect in time, which leads to a significant number of unpredictable consequences.

The paper examined the psychological problems arising from hypothyroidism and hyperthyroidism, namely, analyzed machine learning methods that can calculate the likelihood of developing neurological complications in the setting of these diseases for timely action to eliminate the problem before it occurs.

An analysis of publications describing the ways in which machine learning methods are used has shown that the endocrinological issue is being studied, has prospects for research, and is accompanied by new theories, but these observations try to solve problems that have already arisen or find ways to alleviate the patient's condition without completely eliminating the problem. Therefore, the purpose of the article was to find analytical methods that would predict the likelihood of developing psychological disorders in order to avoid deterioration of patients diagnosed with hypothyroidism or hyperthyroidism.

The paper discusses the symptoms and consequences of hypothyroidism and hyperthyroidism, examples of the implementation of machine learning methods for prognostic purposes in the medical field, proposes the methods studied in this observation based on their common and distinctive characteristics, and analyzes their advantages and disadvantages.

Linear additive convolution was applied to select a more optimal model based on the requirements needed in the outlined task, according to the results of which the "random forest" algorithm is more effective [24, 25].

In the future, selected prediction methods based on medical and psychological indicators will be studied to predict the occurrence of psychological problems due to hypothyroidism and hyperthyroidism in order to introduce certain preventive measures aimed at avoiding neurological complications [26–28].

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2024. No. 2 (28)*

References

1. Pobihun, N.G. (2020), "Research on the impact of physical activity and stress on the thyroid gland in hypothyroidism". *Scientific and Practical Journal*, 3(№ 4 (12)), P. 97–101. available at: https://art-of-medicine.ifnmu.edu.ua/index.php/aom/article/view/402
2. Mubashir Alam, K., Tasnim Ahsan, Urooj Lal, R., Ruqshanda Jabeen, and Saad Farooq. (2017), "Subclinical hypothyroidism: frequency, clinical manifestations, and indications for treatment". *Pakistan Journal of Medical Sciences*, 33(4), P. 818–822. DOI: 10.12669/pjms.334.12921
3. Feldman, A.Z., Shrestha, R.T., & Geneslaw, J.V. (2013), "Neuropsychiatric manifestations of thyroid diseases". *Endocrinology and Metabolism Clinics of North America*, 42(3), P. 453–476. DOI: 10.1016/j.ecl.2013.05.005
4. Almeida, O.P., Alfonso, H., Flicker, L., Hankey, G., Chubb, S.A.P., & Yeap, B.B. (2011), "Thyroid hormones and depression". *The American Journal of Geriatric Psychiatry*, 19(9), P. 763–770. DOI: 10.1097/jgp.0b013e31820dcad5
5. Bunevicius, R., & Prange, A.J. (2010), "Thyroid diseases and mental disorders: cause and effect or only comorbidity?", *Current Opinion in Psychiatry*, 23(4), P. 363–368. DOI: 10.1097/yco.0b013e3283387b50
6. Yarach, D., Kukharska, A., Raevska-Rager, A., & Latska, K. (2012), "Cognitive functions and mood during chronic thyrotropin-suppressive L-thyroxine therapy in patients with differentiated thyroid carcinoma". *Journal of Endocrinological Research*, 35(8), P. 760–765.
7. Demartini, B., Ranieri, R., Masu, A., Selle, V., Scaroni, C., & Gambini, O. (2014), "Depressive symptoms and major depressive disorder in patients with subclinical hypothyroidism". *Journal of Nervous and Mental Disease*, 202(8), P. 603–607. DOI: 10.1097/nmd.0000000000000168
8. Kozhyna, N.M., Tovazhnyanska, O.L., Markova, M.V., Zelenska, K.O., & Kauka, O.I. (2020), "Features of primary hypothyroidism in forcibly displaced persons as a basis for the formation of cognitive and anxiety-depressive disorders". *Problems of Endocrine Pathology*, 73(3), P. 25–32. DOI: https://doi.org/10.21856/j-PEP.2020.3.03
9. Marian, G., Nica, E.A., Ionescu, B.E., & Guinea, D. (2009), "Hyperthyroidism – cause of depression and psychosis: clinical case". Journal of Medicine and Life, 2(4), P. 440–442.
10. Dabldey, A.R., & Sippel, R.S. (2020), "Hyperthyroidism". *Gland Surgery*, 9(1), P. 124-135. DOI: 10.21037/gs.2019.11.01
11. Soiri, I.N., & Reidpat, D.D. (2013), "Health forecasting review", *Environmental Health and Preventive Medicine*, №18, P. 1–9. https://doi.org/10.1007/s12199-012-0294-6
12. Armstrong, J.S. (2001), *"Principles of forecasting: A handbook for researchers and practitioners"*. Norwell: Kluwer Academic Publishers. 458 p.
13. Savchuk, O. (2021), "Application of machine learning in clinical psychology". available at: https://ojs.tdmu.edu.ua/index.php/kl-stomat/article/download/6147/5624/21744
14. Hodovanyets, O.I. & Rozhko, M.M. (2015), "Features of the formation of the dental arch system in children with diffuse non-toxic goiter", *Bulletin of Biology and Medicine Issues*, Vol. 2, Issue 2(119), P. 37–39.
15. Zelinska, N.B., Tereshchenko, A.V., & Rudenko, N.G. (2013), "The state of providing specialized assistance to children with endocrine pathology in Ukraine in 2012 and prospects for its development". *Ukrainian Journal of Pediatric Endocrinology*, No 3, P. 31–39.
16. Lytvynenko, O. (2021), "Innovative approaches to processing psychological data using machine learning". available at: https://openarchive.nure.ua/server/api/core/bitstreams/86aa5c34-6f0f-44a4-ac51-76e7d191e085/content
17. Livingston, E.H. (2019), "Subclinical hypothyroidism". *JAMA*, 322(2), 180 p. DOI: 10.1001/jama.2019.9508
18. Kyslyi, O. (2021), "Using machine learning methods in psychological research". available at: https://ami-ejournal.cdu.edu.ua/article/view/4158/4438
19. Zbarazhskyi, M. (2021), "Analysis of psychological data using machine learning techniques". available at: https://ojs.tdmu.edu.ua/index.php/visnyk-nauk-dos/article/view/8460/7880
20. Ivanov, I. (2021), "Advanced machine learning techniques in psychology". available at: https://ela.kpi.ua/items/20f948bd-5b8a-420e-a86e-70d86be50866
21. Kyslyi, O. (2020), "Using machine learning methods in psychological research. Artificial Intelligence Methods". available at: https://ami-ejournal.cdu.edu.ua/article/view/4158/4438
22. Petrov, A. (2021), "Applications of machine learning in psychological studies". available at: https://ela.kpi.ua/server/api/core/bitstreams/17dfafd7-9874-4b51-b865-f20ea63e5076/content
23. Cutler, A., & Zhao, G. (2001), "*PERT – Perfect Random Tree Ensembles*". Computing Science and Statistics, № 33, P. 490–497.

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №2 (28)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

24. Babak, V.P., Biletskyi, A.Ya., Prystavka, O.P., & Prystavka, P.O. (2001), *"Statistical data processing"*. Kyiv: MIVVTS. 388 p.
25. Breiman, L. (2001), *"Random Forests"*. Machine Learning, P. 45.
26. Mochurad, L. & Ilkiv, A. (2022), "Advanced method of medical classification using parallelization algorithms". *Computer Systems and Information Technologies*, (1), P. 23–31. DOI: 10.31891/CSIT-2022-1-3
27. Ittermann, T., Fiolka, H., Baumeister, S.E., Appel, K., & Graabe, H.J. (2015), "Diagnosed thyroid diseases associated with depression and anxiety". *Social Psychiatry and Psychiatric Epidemiology*, 50(9), P. 1417–1425. DOI: 10.1007/s00127-015-1043-0
28. Martino, J.P. (1972), *"Forecasting the progress of technologies"*. New York, New York: Gordon and Breach Science Publishers. № 2. 15 p.

Надійшла (Received) 08.05.2024

Відомості про авторів / About the Authors

Гулієв Нурал Бахадур огли – Харківський національний університет радіоелектроніки, аспірант, Харків, Україна; e-mail: nural.huliiev@nure.ua; ORCID ID: https://orcid.org/0000-0003-2123-0377

Huliiev Nural Bahadur ohli – Kharkiv National University of Radio Electronics, PhD Student, Kharkiv, Ukraine.

ВИБІР МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ РОЗВИТКУ ПСИХОЛОГІЧНИХ РОЗЛАДІВ У ЛЮДЕЙ ІЗ ГІПОТИРЕОЗОМ ТА ГІПЕРТИРЕОЗОМ

Предметом дослідження в статті є ендокринологічні захворювання, а саме: аналіз ускладнень у людей з гіпотиреозом та гіпертиреозом. Відомо, що ці хвороби виникають безсимптомно або можуть бути наслідками інших захворювань, через що люди не підозрюють, на що саме хворіють. Пізніше хвороби зазвичай спричиняють ускладнення в організмі, найнебезпечнішими з яких є психологічні розлади: депресія, маніакальність, агресивність тощо. Тому **метою роботи** є розроблення методів прогнозування виникнення неврологічних погіршень організму в людей, у яких вже виявлено ендокринологічні захворювання. У статті розв'язувалися **завдання** вибору кращих моделей прогнозування виникнення психологічних розладів у пацієнтів з ендокринологічними проблемами. Аналізувалися методи машинного навчання, поширені в медичній галузі, та обирався один із них, який найбільш ефективно вирішує всі поставлені завдання. У виборі критеріїв узято до уваги потенційні проблеми з медичними та психологічними показниками. Упроваджувався **метод** лінійної адитивної згортки для вибору найкращих за результатами альтернатив, із принципом Парето, спрямованим на вилучення непідхожих альтернатив через те, що всі ознаки мають менші показники, ніж в інших варіантах. Для експерименту всі ознаки конвертувалися в кількісні для підрахунку значень згортки. Критерії оцінки наведені в роботі. Досягнуто таких **результатів**: моделлю прогнозування в подальшому дослідженні окресленого завдання буде випадковий ліс. **Висновки**: досліджено методи прогнозування та обрано більш оптимальну модель за допомогою лінійної адитивної згортки, а саме алгоритм "випадковий ліс", розглянуто переваги й недоліки зазначеної моделі. Більш детальний аналіз її розроблення буде запропоновано в наступних статтях. Надано математичний опис обраного методу прогнозування, що містить потенційні способи реалізації та кроки побудови алгоритму одного із цих способів.

Ключові слова: гіпотиреоз; гіпертиреоз; психологічні розлади; прогнозування; лінійна адитивна згортка; принцип Парето; алгоритм "випадковий ліс"; дерево рішень; індекс *Gini*.

Бібліографічні описи / Bibliographic descriptions

Гулієв Н. Б. Вибір моделей машинного навчання для прогнозування розвитку психологічних розладів у людей із гіпотиреозом та гіпертиреозом. *Сучасний стан наукових досліджень та технологій в промисловості*. 2024. № 2 (28). С. 76–85. DOI: https://doi.org/10.30837/2522-9818.2024.2.076

Huliiev, N. (2024), "Choice of machine learning models for predicting the development of psychological disorders in people with hypothireosis and hyperthireosis", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (28), P. 76–85. DOI: https://doi.org/10.30837/2522-9818.2024.2.076