Classification and Prediction of Diabetes Disease using Decision Tree Method

Tetiana Dudkina[a], Ievgen Meniailov[a], Kseniia Bazilevych[a], Serhii Krivtsov[a] and Anton Tkachenko[b]

[a] *National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine*
[b] *Kharkiv National Medical University, Kharkiv, Ukraine*

Abstract

Digitalization in medicine has become one of the largest gaps in almost all healthcare systems in the world. Diabetes remains one of the pressing health problems. According to World Health Organization, the number of people with diabetes increased from 108 million in 1980 to 422 million in 2014. This research is devoted to solving the problem of classifying patients with diabetes and diagnosing this disease. To solve the problem, a machine learning model was built based on a decision tree method. To develop the model, an open database of patients with diabetes, consisting of 768 patients, was used. On the foundation of the constructed model, a software package in the Python language has been developed.

Keywords [1]

Machine Learning, Diabetes, Classification, Decision Tree, Prediction.

1. Introduction

The world pandemic of the new coronavirus has changed the usual way of life and approaches to solve medical tasks [1].

Right now, digitalization in medicine has become one of the largest gaps in almost all healthcare systems in the world. As practice shows, digital technologies can significantly improve the quality of healthcare [2]. For example, modern models of the spread of infectious diseases [3], such as HIV [4], tuberculous [5], hepatitis B [6], influenza and ARVI [7], syphilis [8], and others, make it possible to predict the incidence and develop effective preventive measures to reduce the incidence. The development of management systems [9-11] for medical institutions and medical insurance systems [12] allows automating decision making. Information technologies help medical staff during surgeries [13-15]. Automated training systems for medical personnel allow timely updating of their knowledge [16-18]. Modern techniques of medical images analysis [19-20] and methods for diagnosing common diseases such as cancer [21-22] or heart disease [23] can detect diseases at an early stage.

Diabetes remains one of the pressing health problems. According to World Health Organization, the number of people with diabetes increased from 108 million in 1980 to 422 million in 2014 [24]. The global prevalence of diabetes among people over 18 years of age increased from 4.7% in 1980 to 8.5% in 2014 [25]. Premature mortality from diabetes increased by 5% between 2000 and 2016 [26]. Some scientists have linked cases of childhood diabetes with COVID-19 [27].

Diabetes mellitus is a chronic disease of the endocrine system, which is caused by a violation of insulin synthesis and an increase in blood sugar. The disease can lead to the development of a number of serious deficiencies. There are 2 main types of diabetes mellitus: types I and II, as well as gestational diabetes in pregnant women and symptomatic diabetes. Let's consider 2 main types, from which more and more people are suffering from all over the globe every day. Diabetes mellitus type I is more

IT&AS 2021: Symposium on Information Technologies & Applied Sciences, March, 5, 2021, Bratislava, Slovakia
EMAIL: dudkinatetiana@gmail.com (T. Dudkina); evgenii.menyailov@gmail.com (I. Meniailov); ksenia.bazilevich@gmail.com (K. Bazilevych); krivtsovpro@gmail.com (S. Krivtsov); antontkachenko555@gmail.com (A. Tkachenko).
ORCID: 0000-0001-6309-2836 (T. Dudkina); 0000-0002-9440-8378 (I. Meniailov); 0000-0001-5332-9545 (K. Bazilevych); 0000-0001-5214-0927 (S. Krivtsov); 0000-0002-1029-1636 (A. Tkachenko).

common in patients under the age of 30, more often it develops due to the fact that the pancreas begins to work worse against the background of a viral infection or the action of toxins. With this type of diabetes, the body is unable to produce insulin, thus, when diagnosed with type I diabetes, the patient becomes insulin dependent throughout his life. Diabetes mellitus type II - occurs due to insulin resistance, of which they get sick more often. Older people are more prone to it, because sugar tolerance decreases over the years. There are a number of factors that increase your risk of developing type II diabetes. [28].

The symptoms depend on how long the person is sick, on the severity of the disease and the patient's personal immunity. Someone may have a vivid clinical picture right away, while someone may have a barely noticeable clinic or, even worse, be absent. Diabetes can be diagnosed using a variety of diagnostics. The main method for diagnosing diabetes mellitus is laboratory tests of urine and blood for glucose levels. In some cases, the doctor may prescribe an ultrasound of the kidneys, an EEG of the brain, etc. People at risk should carefully monitor their blood glucose and blood pressure levels.

An important challenge in the fight against diabetes is the classification of patients and the diagnosis of the disease. To solve this problem, it is advisable to use a machine learning apparatus. In modern science, several models have been implemented that make it possible to diagnose diabetes according to specified parameters.

Sisodia S. and Sisodia D.S. have made prediction model of Diabetes using naïve Bayes algorithm with accuracy 76.3% [29]. Naveen K. has made classification model of Diabetes using SVM algorithm and data of glucose and blood pressure [30].

These and other analyzed researches show the limitations of the factors used to train the model, which leads to a decrease in the classification accuracy.

The **aim of the research** is to build a model that allows classifying a person's condition in relation to the incidence of diabetes using machine learning methods.

2. Materials and Methods

To solve the problem, we use the decision trees method [31-32]. The decision tree is a sequential hierarchical structure and includes: branches with attributes on which the result depends - the objective function; nodes - random vertices in which possible scenarios for the development of events are determined; leaf (leaf) nodes with objective function values represent the final results of choosing a specific attribute value and combine several objects. Decision trees are divided into two types by the type of predicted indicator: classification trees and regression trees. When developing a system for establishing a diagnosis, it is advisable to use classification trees, since they are used research on certain attributes, namely, to attribute objects (symptoms) from a previously known class (a certain disease). Decision trees divide data into groups, resulting in a hierarchy of "if ... then ..." operators that classifies data.

Let's define an objective function in order to divide the nodes into informative functions. Each partition where we maximize the increment is:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \quad (1)$$

where f is attribute by which splitting is performed; D_p and D_j are parent and j-th child nodes; I is a measure of heterogeneity; N_p is the total number of samples in the parent node; N_j is the number of samples in the j-th child node.

For simplicity and to reduce the combinatorial search space, we implement binary decision trees. In our case, child nodes D_{left} and D_{right} are:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \quad (2)$$

where f is attribute by which splitting is performed; D_p and D_j are datasets of parent and j-th child nodes; I is a measure of heterogeneity; N_p is total number of samples in parent node; D_{left} and D_{right} are child

nodes; N_{left} and N_{right} are numbers of patterns in left and right child nodes; N_j is number of samples in j-th child node.

Determination of entropy for all non-empty classes $p(i|t) \neq 0^2$:

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) log_2 p(i|t) \quad (3)$$

where $p(i|t)$ is fraction of samples that belongs to class and single node *t*.

So, the entropy is 0 if all samples in a node belong to the same class, and the entropy is maximal if we have a uniform distribution of classes.

The Gini measure of heterogeneity [33] can be perceived as a criterion that minimizes the likelihood of misclassification:

$$I_G(t) = \sum_{i=1}^{c} p(i|t)\big(1 - p(i|t)\big) = 1 - \sum_{i=1}^{c} p(i|t)^2 \quad (4)$$

where $p(i|t)$ is fraction of samples that belongs to a class and a single node *t*; $L_G(t)$ is Gini measure of heterogeneity.

Another measure of heterogeneity is classification error:

$$I_\varepsilon(t) = 1 - \max\{p(i|t)\} \quad (5)$$

where $p(i|t)$ is fraction of samples that belongs to a class and a single node *t*; $I_\varepsilon(t)$ is classification error.

This criterion is suitable for tree pruning, but is not recommended for tree growth because it is less sensitive to changes in the capabilities of the classes in the nodes.

3. Implementation and results

In order to build a decision tree, you need certain data. The Pima Indians Diabetes DataBase was used to test the diabetes diagnostic model. Database has 768 instances and 9 attributes for individual patients (Table 1).

Table 1

Full list of parameters

Parameter name	Description	Data type
Pregnancies	number	decimal
PG Concentration	count	integer
Diastolic BP	count	integer
Tri Fold Thick	count	integer
Serum Ins	count	integer
BMI	count	integer
DP Function	count	integer
Age	years	decimal
Diabetes	present or not	0/1

The data distribution is shown in Figure 1.

Since data is the most important factor for effective work of the classifier, it is important to know which features from the dataset have the greatest impact on the classification, and which ones, on the contrary, have no effect at all. With this information, you can manipulate what data to input. Thus, you can increase the value of the system, because at this stage, after receiving the results, a doctor's consultation and more accurate diagnosis are still needed. There are certain signs in this dataset that the user will not be able to find out for himself, however, there is a possibility that the value of the system will fall to him if he first learns these signs from a doctor.

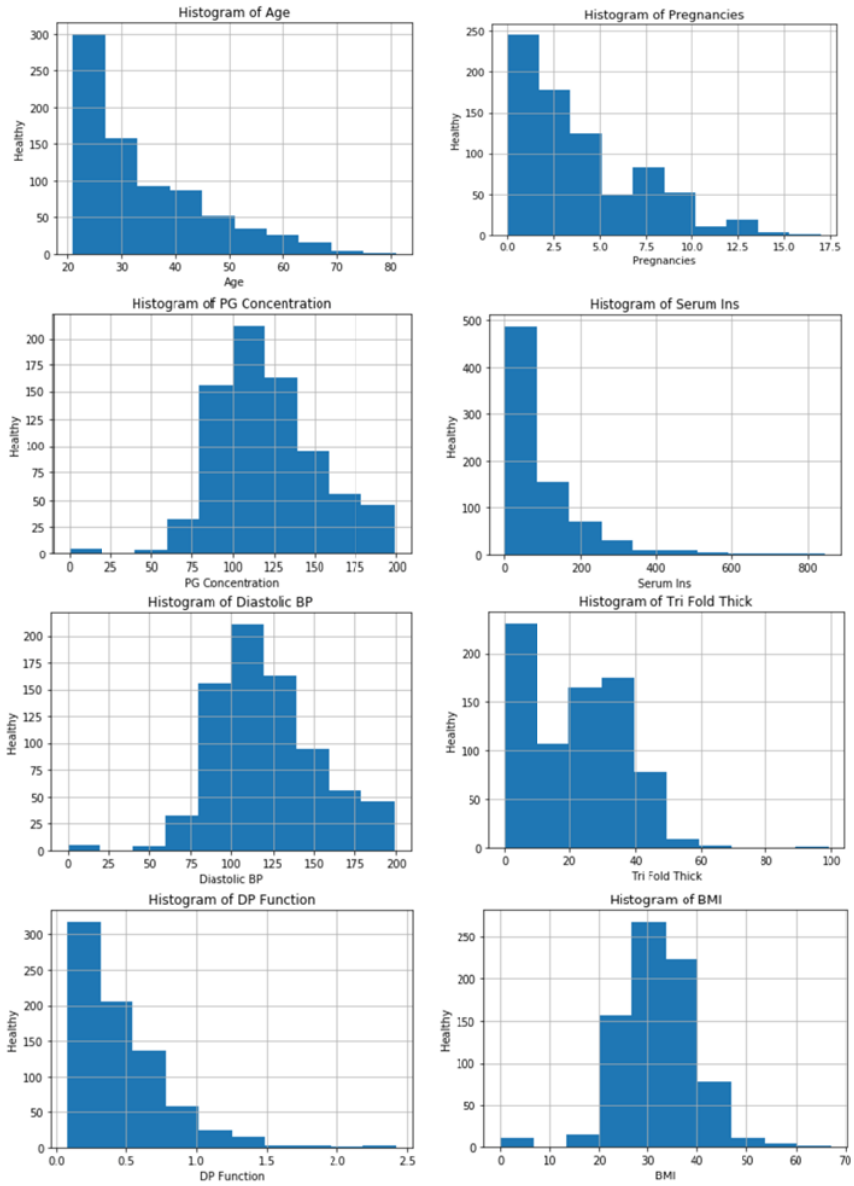

Figure 1: Data distribution

As we can see from the plot, there are some outliers in some of the columns.

We can see that there are 0 values for blood pressure. So, we assume that it is mistake data. Observing the data, we see 35 samples, where the value is 0. Even after fasting, your glucose level will not go below zero. Therefore, zero is a misread. Before further use of the selection, remove the lines with “BloodPressure”, “BMI” and “Glucose” equal to zero.

The Spyder development environment was used to write the code to build the decision tree. In order to read our data from the table, the Pandas library was used. The Scikit-learn machine learning library was also used. To implement a decision tree, you need to import the required Python packages. Then we upload our database.

The next step is to split this data into two parts - training data and testing data. Next, you need to train the model using the DecisionTreeClassifier class (Scikit-learn library). Next, we make a forecast, and we also need to get an accuracy estimate, a classification report and an error matrix. The final step is to render our decision tree [34]. The color of the nodes is used to highlight the class that has the most in each node and to convey the names of the classes and traits so that the tree is correctly marked up.

For the first experiment, the data was split as follows: 70% for training and 30% for testing. The results are shown in Figures 2 and 3.

```
Confusion Matrix:
[[112  34]
 [ 45  40]]
('Classification Report:',)
              precision    recall  f1-score
support

     Healthy       0.71      0.77      0.74
146
        Sick       0.54      0.47      0.50
85

   micro avg       0.66      0.66      0.66
231
   macro avg       0.63      0.62      0.62
231
weighted avg       0.65      0.66      0.65
231

('Accuracy:', 0.658008658008658)
```

Figure 2: Results of experiment 1 (70% for training, 30% for testing).

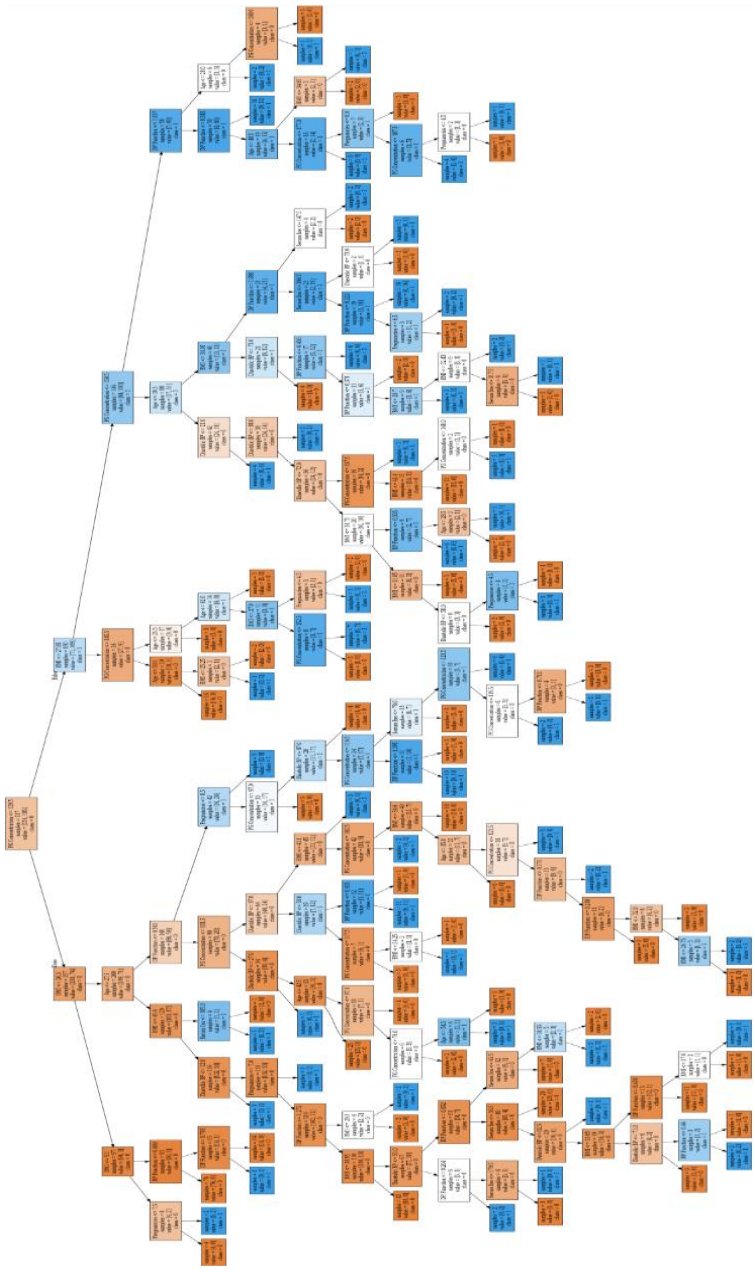

Figure 3: Visualization of experiment 1 (70% for training, 30% for testing).

For the second experiment, the data was split as follows: 50% for training and 50% for testing. The results are shown in Figures 4 and 5.

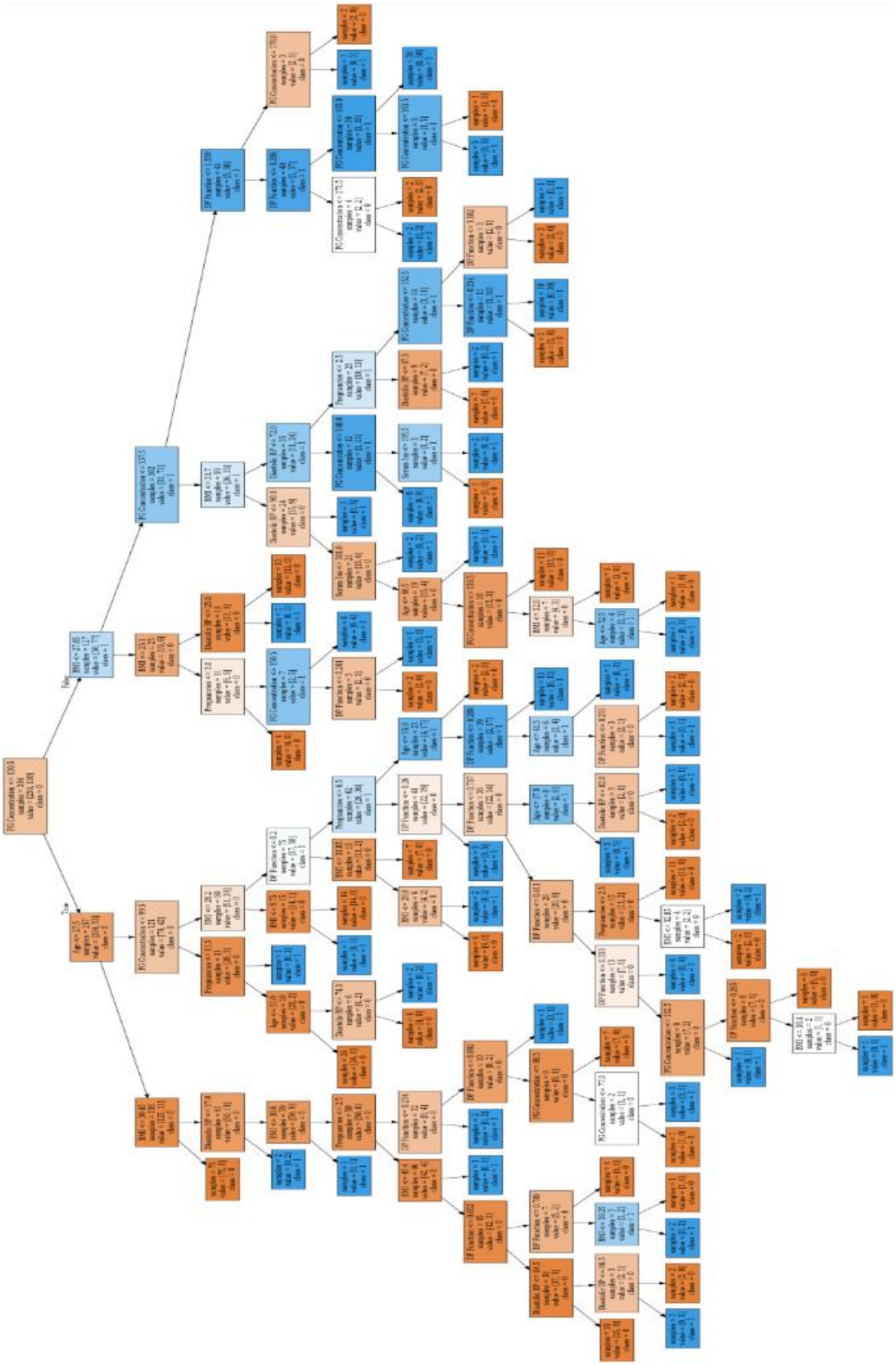

Figure 4: Visualization of experiment 2 (50% for training, 50% for testing).

```
Confusion Matrix:
[[195  51]
 [ 59  79]]
('Classification Report:',)
              precision    recall  f1-score
support

     Healthy       0.77      0.79      0.78
246
        Sick       0.61      0.57      0.59
138

   micro avg       0.71      0.71      0.71
384
   macro avg       0.69      0.68      0.68
384
weighted avg       0.71      0.71      0.71
384

('Accuracy:', 0.7135416666666666)
```

Figure 5: Results of experiment 2 (50% for training, 50% for testing).

For the third experiment, the data was split as follows: 30% for training and 70% for testing. The results are shown in Figures 6 and 7.

```
Confusion Matrix:
[[251 103]
 [ 95  89]]
('Classification Report:',)
              precision    recall  f1-score
support

     Healthy       0.73      0.71      0.72
354
        Sick       0.46      0.48      0.47
184

   micro avg       0.63      0.63      0.63
538
   macro avg       0.59      0.60      0.60
538
weighted avg       0.64      0.63      0.63
538

('Accuracy:', 0.6319702602230484)
```

Figure 6: Results of experiment 3 (30% for training, 70% for testing).

4. Conclusions

Overall, it can be said that decision tree analysis is a predictive modeling tool that can be applied in many areas. Decision trees can be built using an algorithmic approach that can partition the dataset in different ways depending on conditions.

After the work done, we can conclude that the more data is allocated for training the model, the better the accuracy estimate we get. In our case, the best option is to split the data by 50% for training the model and 50% for testing, since the accuracy of this option is 0.71.

After analyzing the constructed diagnostic model, the following advantages can be identified: fast learning process; generation of rules in areas where it is difficult for an expert to formalize his knowledge; intuitive classification model; high prediction accuracy, comparable to other methods of data analysis (statistics, neural networks); construction of nonparametric models.

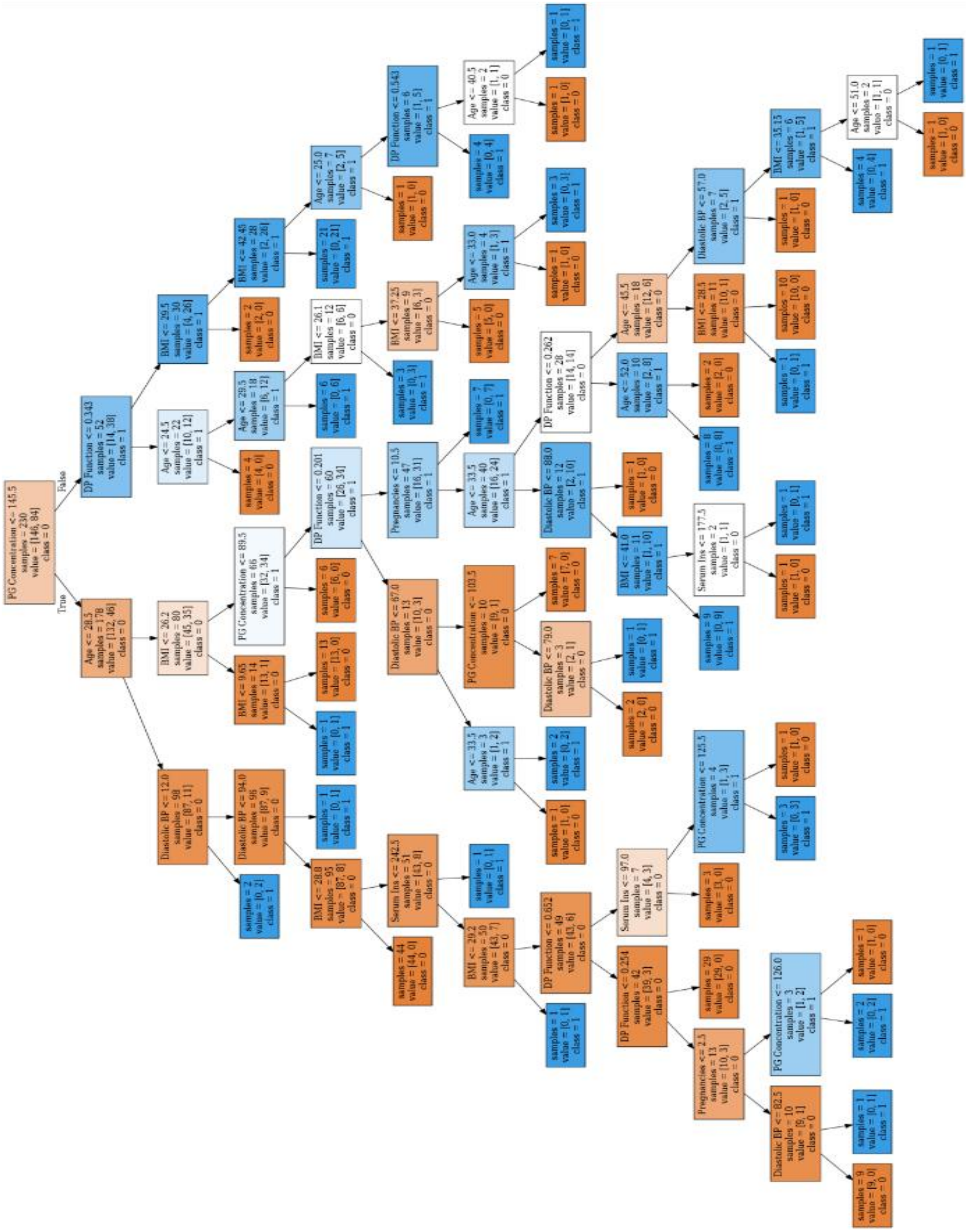

Figure 6: Visualization of experiment 3 (30% for training, 70% for testing).

5. Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic "Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management" [35].

References

[1] G. Pascarella, et al. COVID-19 diagnosis and management: a comprehensive review, Journal of International Medicine 288(2) (2020) 192-206.

[2] M. Mazorchuck, et. al. Web-Application Development for Tasks of Prediction in Medical Domain, International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT) (2018) 5-8.

[3] Yu. Polyvianna, et. al. Computer Aided System of Time Series Analysis Methods for Forecasting the Epidemics Outbreaks, International Conference on the Experience of Designing and Application of CAD Systems (2019) pp. 7.1-7.4. doi: 10.1109/CADSM.2019.8779344

[4] D. Chumachenko, T. Chumachenko, Intelligent Agent-Based Simulation of HIV Epidemic Process, Advances in Intelligent Systems and Computing 1020 (2019) 175-188. doi: 10.1007/978-3-030-26474-1_13

[5] D. Chumachenko, et. al. On-Line Data Processing, Simulation and Forecasting of the Coronavirus Disease (COVID-19) Propagation in Ukraine Based on Machine Learning Approach, Communications in Computer and Information Science 1158 (2020) 372-382. doi: 10.1007/978-3-030-61656-4_25

[6] D. Chumachenko, On Intelligent Multiagent Approach to Viral Hepatitis B Epidemic Processes Simulation, International Conference on Data Stream Mining and Processing (2018) 415-419.

[7] T. Banirostam, M. N. Fesharaki, Modeling and Simulation of Influenza with Biological Agent: A New Approch for Increasing System Robustness, Fifth Asia Modelling Symposium, Kuala Lumpur (2011) 13-17.

[8] Chumachenko D. , et. al. Development of an intelligent agent-based model of the epidemic process of syphilis, International Scientific and Technical Conference on Computer Sciences and Information Technologies (2019) 42-45.

[9] Dotsenko N. , et. al. Modeling of the process of critical competencies management in the multi-project environment, International Scientific and Technical Conference on Computer Sciences and Information Technologies 3 (2019) 89-93.

[10] Dotsenko N. , et. al. Project-oriented management of adaptive teams' formation resources in multi-project environment, CEUR Workshop Proceedings 2353 (2019) 911-920.

[11] Dotsenko N. , et. al. Modeling of the processes of stakeholder involvement in command management in a multi-project environment, International Scientific and Technical Conference on Computer Sciences and Information Technologies 1 (2018) 29-33.

[12] Bazilevych K. , et al. Stochastic modelling of cash flow for personal insurance fund using the cloud data storage, International Journal of Computing 17 (3) (2018) 153-162.

[13] Bohdanov S. , et. al., Forecasting of salmonellosis epidemic proces in Ukraine using autoregressive integrated moving average model, Przeglad epidemiologiczny 74 (2) (2020) 346–354.

[14] Chumachenko D. , Chumachenko K. , Yakovlev S. , Intelligent simulation of network worm propagation using the code red as an example, Telecommunications and Radio Engineering 78 (5) (2019) 443-464. doi: 10.1615/TelecomRadEng.v78.i5.60

[15] Chumachenko D. , Yakovlev S. , On intelligent agent-based simulation of network worms propagation, 2019 15th International Conference on the Experience of Designing and Application of CAD Systems (2019) 11-15. doi: 10.1109/CADSM.2019.8779342

[16] P. Piletskiy, et. al. Development and Analysis of Intelligent Recommendation System Using Machine Learning Approach, Advances in Intelligent Systems and Computing 1113 (2020) 186-197.

[17] A. Herasymova, et. al., Development of intelligent information technology of computer processing of pedagogical tests open tasks based on machine learning approach, CEUR, 2631 (2020) 121-131.

[18] D. Chumachenko, et. al. Intelligent expert system of knowledge examination of medical staff regarding infections associated with the provision of medical care, CEUR, 2386 (2019) 321-330.

[19] V. P. Mashtalir, S. V. Yakovlev, Point-set methods of clusterization of standard information, Cybernetics and Systems Analysis 37 (3) (2001) 295-307.

[20] V. P. Mashtalir, et al. Group structures on quotient sets in classification problems, Cybernetics and Systems Analysis 50 (4) (2014) 507-518.
[21] I. Meniailov, et. al. Using the K-means method for diagnosing cancer stage using the Pandas library, CEUR, 2386 (2019) 107-116.
[22] D. Chumachenko, et. al. On agent-based approach to influenza and acute respiratory virus infection simulation, 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (2018) 192-196. doi: 10.1109/TCSET.2018.8336184
[23] K. Bazilevych, et.al. Determining the Probability of Heart Disease using Data Mining Methods. CEUR, 2488 (2019) 1-12.
[24] G. Valenti, G, Tamma, History of Diabetes Insipidus, Giornale italiano di nefrologia 33 (2016) 66:33.S66.1.
[25] N. Sarwar, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies, Lancet 375 (9733) (2014) 2215-2222.
[26] D. Chumachenko, et. al. On Intelligent Decision Making in Multiagent Systems in Conditions of Uncertainty, Proceedings of 2019 11th International Scientific and Practical Conference on Electronics and Information Technologies (2019) 150-154. doi: 10.1109/ELIT.2019.8892307
[27] A. Hussain, B. Bhowmik, N. C. do Vale Moreira, COVID-19 and diabetes: Knowledge in progress, Diabetes Research and Clinical Practice 162 (2020) 108142.
[28] D.G. Bichet, Genetics and diagnosis of central diabetes insipidus, Annales d'Endocrinologie 73 (2) (2012) 117-127.
[29] V. Yesina, et. al., Method of Data Openness Estimation Based on User-Experience in Infocommunication Systems of Municipal Enterprises, International Scientific-Practical Conference on Problems of Infocommunications Science and Technology (2019) 171–176. doi: 10.1109/INFOCOMMST.2018.8631897
[30] K.G. Naveen, et. al., Prediction of diabetes using Machine Learning classification algorithms, International journal of scientific and technology research 9 (1) (2020) 1805-1808.
[31] A. Albu, From logical inference to decision trees in medical diagnosis, Proceedings of 2017 E-Health and Bioengineering Conference (2017) 65-68. doi: 10.1109/EHB.2017.7995362
[32] M.D.A. Praveena, J. S. Krupa, S. SaiPreethi, Statistical Analysis Of Medical Appointments Using Decision Tree, Conference on Science Technology Engineering and Mathematics (2019) 59-64. doi: 10.1109/ICONSTEM.2019.8918766
[33] D. Chumachenko, O. Sokolov, S. Yakovlev, Fuzzy recurrent mappings in multiagent simulation of population dynamics, International Journal of Computing 19 (2) (2020) 290-297.
[34] S. N. Gerasin, et. al., Set coverings and tolerance relations, Cybernetics and Systems Analysis 44 (3) (2008) 333-340.
[35] Yakovlev S. , et. al., The concept of developing a decision support system for the epidemic morbidity control, CEUR, 2753 (2020) 265–274.