Investigation of Machine Learning Methods for Stroke Prediction

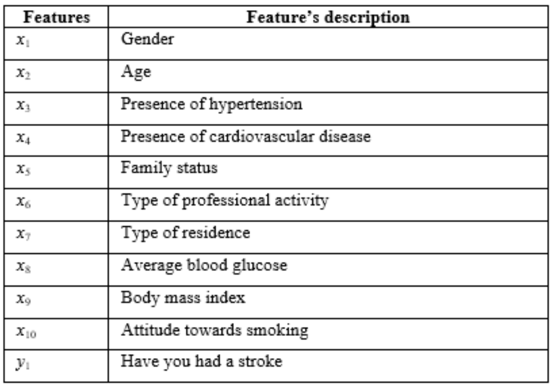
V.V.Mokshin, A.R.Faskhutdinova,D.N.Grigorieva,B.A.Garafutdinov

In the medical field every day work with a large amount of data. This could be input about patients' symptoms or collection of tests.

All data is analyzed and a data table is created. Relationships between data are identified, models are built on their basis.

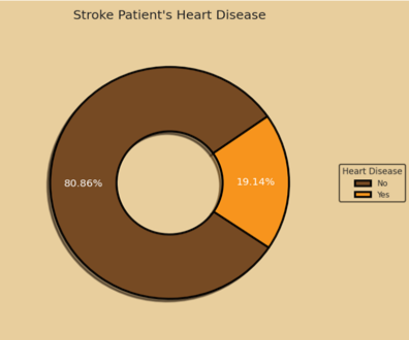
The output will be the correctness of the diagnoses based on the symptoms or the accuracy of the analysis. Data analysis methods will help identify people susceptible to the disease and classify what factors affect it. For this, machine learning is used.

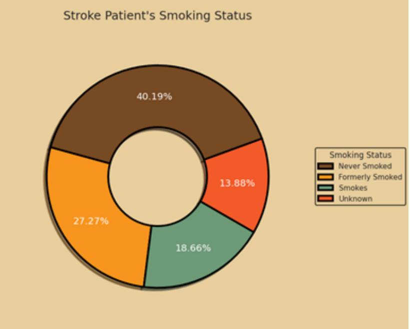
TABLE I. FEATURES FOR ANALYSIS

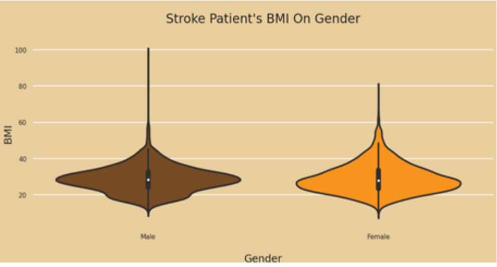


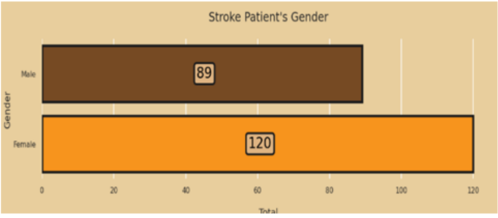
Features	Feature's description
x_1	Gender
x_2	Age
x_3	Presence of hypertension
x_4	Presence of cardiovascular disease
x_5	Family status
x_6	Type of professional activity
x_7	Type of residence
x_8	Average blood glucose
x_9	Body mass index
x_{10}	Attitude towards smoking
y_i	Have you had a stroke

The aim of this work is to develop a model that can quickly and accurately identify the risk of stroke based on a small number of input parameters. For the most accurate and reliable result, a large number of implementation methods are considered and compared. During the comparison, the method with the highest execution accuracy was selected.

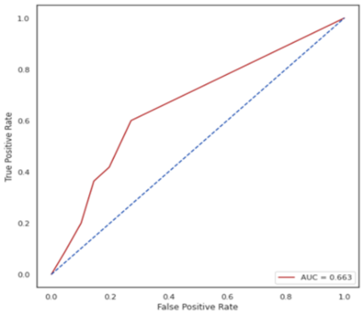




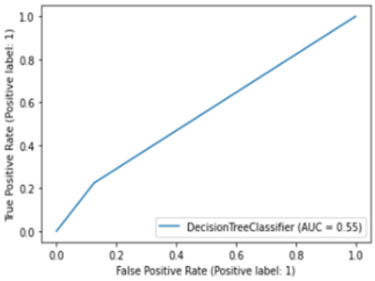




$$Q_j = \sum_{i=1}^{n_j} \frac{1}{D^2(x, a_{ij})},$$



$$\frac{b}{\square\vec{\omega}\square},$$

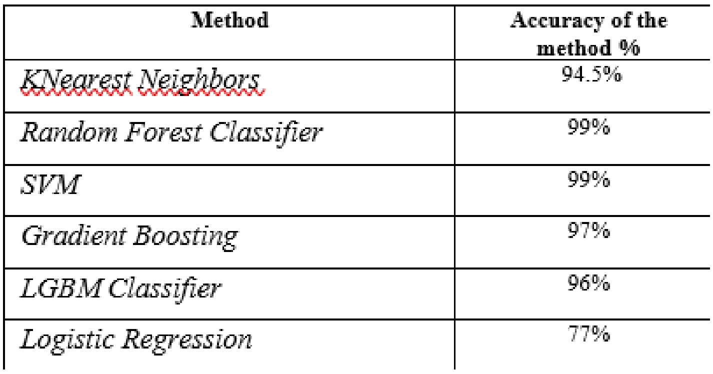


The problem of learning on unbalanced data is a fairly common topic for research in recent years. The presence of this proble was taken into account when creating the neural network architecture. The effectiveness of using artificial intelligence in assessing the risk of developing cardiovascular diseases was substantiated in.

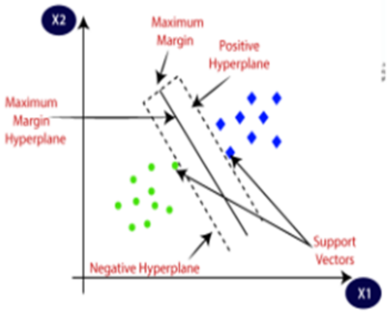
$$\langle \omega^{\rightarrow}, x^{\rightarrow} \rangle,\ b = 0$$

Logistic function formula:

$$f(x) = \frac{1}{1+e^{-x}}$$



Method	Accuracy of the method %
KNearest Neighbors	94.5%
Random Forest Classifier	99%
SVM	99%
Gradient Boosting	97%
LGBM Classifier	96%
Logistic Regression	77%



[1] **KAZAN NATIONAL RESEARCH TECHNICAL UNIVERSITY NAMED AFTER A. N. TUPOLEV - KAI**
10 K.Marx St., Kazan,
Tatarstan 420111, Russia (www.kai.ru)