УДК 004.9:616.831-005

**Є. Ю. Гонтковський**
**О. М. Козачко**

# ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПЕРЕДБАЧЕННЯ ХВОРИХ НА ІНСУЛЬТ

Вінницький національний технічний університет

**Анотація**

*В роботі розроблено технологію передбачення хворих на інсульт. Для розробки технології, було використано готовий набір даних, що включає с себе дані пацієнтів. Виконано прогнозування даних використовуючи моделі Logistic Regression, KNearest Neighbors, Decision Tree Classifier, Random Forest Classifier, Ada Boost, Support Vector Machine (SVM), XGBoost, Cat Boost.*

**Ключові слова**: Python, інсульт, розвідувальний аналіз, хвороба

**Abstract**

*The developed technology focuses on predicting stroke cases. To create this technology, an existing dataset containing patient information was utilized. Predictions were made using the following models: Logistic Regression, K-Nearest Neighbors, Decision Tree Classifier, Random Forest Classifier, Ada Boost, Support Vector Machine (SVM), XGBoost, and Cat Boost..*

**Key words**: Python, stroke, intellectual analysis, disease.

## Вступ

Інформаційна технологія передбачення хворих на інсульт займає важливе місце серед актуальних проблем сучасної медицини та інформаційних технологій. З урахуванням того, що інсульт є однією з найпоширеніших і найнебезпечніших неврологічних патологій, його своєчасне виявлення та передбачення стає важливим завданням для забезпечення ефективної медичної допомоги та попередження негативних наслідків [1].

Актуальність теми обумовлена необхідністю вдосконалення методів діагностики та прогнозування інсульту, адже час відкриття та невідкладного лікування має ключове значення для підвищення шансів на повне відновлення хворого. Інформаційні технології, такі як штучний інтелект, машинне навчання та аналіз великих обсягів даних, надають унікальні можливості для створення точних та ефективних систем передбачення ризику інсульту.

## Постановка задачі

Метою роботи є розроблення інформаційної технології для аналізу та передбачення стану хворих на гепатит з використанням методів машинного навчання.

Для досягнення цієї мети необхідно вирішити наступні завдання:

- повести огляд існуючих систем;
- підготувати дані для подальшої роботи;
- провести розвідувальних аналіз даних;
- побудувати моделі та виконати прогнозування;
- оцінити результати роботи моделей.

## Результати дослідження

Даними для аналізу та передбачення було обрано датасет «Kaggle Stroke Prediction Dataset» у середовищі Kaggle [3]. Приклад даних з цього датасету показано на рисунку 1.

[3]:

| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 9046 | Male | 67.0 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked | 1 |
| 1 | 51676 | Female | 61.0 | 0 | 0 | Yes | Self-employed | Rural | 202.21 | NaN | never smoked | 1 |
| 2 | 31112 | Male | 80.0 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked | 1 |
| 3 | 60182 | Female | 49.0 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 |
| 4 | 1665 | Female | 79.0 | 1 | 0 | Yes | Self-employed | Rural | 174.12 | 24.0 | never smoked | 1 |

Рис. 1 – Перші 5 стовпців датасету

На рисунку 2 показано кореляційну матрицю датасету.

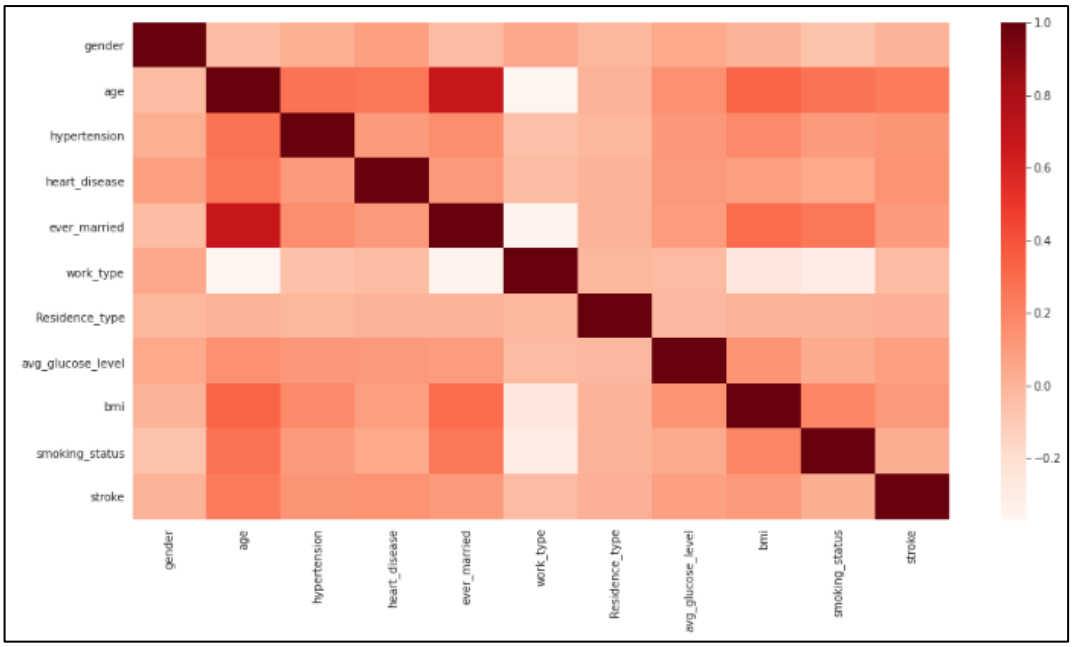


Рис. 2 – Кореляційна матриця датасету

Перед побудовою моделей було використано пакет SMOTE для збалансування датасету. Для виправлення пропущених значень у стовпчиках BMI ми використаємо техніку імплікації на основі KNN (рис. 3-4).

.

```
en_df_imputed = en_df
imputer = KNNImputer(n_neighbors=4, weights="uniform")
imputer.fit_transform(en_df_imputed)
```

```
array([[  1.,  88.,   0., ..., 239.,   1.,   1.],
       [  0.,  82.,   0., ..., 418.,   2.,   1.],
       [  1., 101.,   0., ..., 198.,   2.,   1.],
       ...,
       [  0.,  56.,   0., ..., 179.,   2.,   0.],
       [  1.,  72.,   0., ..., 129.,   1.,   0.],
       [  0.,  65.,   0., ..., 135.,   0.,   0.]])
```

+ Code + Markdown

```
en_df_imputed.isnull().sum()
```

```
gender              0
age                 0
hypertension        0
heart_disease       0
ever_married        0
work_type           0
Residence_type      0
avg_glucose_level   0
bmi                 0
smoking_status      0
stroke              0
dtype: int64
```

Рисунок 3 – KNNImputer для заміни відсутніх значень

```
from imblearn.over_sampling import SMOTE
X , y = en_df_imputed[features],en_df_imputed["stroke"]
x_train,x_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=23)
sm = SMOTE()
X_res, y_res = sm.fit_resample(x_train,y_train)

print("Before OverSampling, counts of label '1': {}".format(sum(y==1)))
print("Before OverSampling, counts of label '0': {} \n".format(sum(y==0)))

print('After OverSampling, the shape of train_X: {}'.format(X_res.shape))
print('After OverSampling, the shape of train_y: {} \n'.format(y_res.shape))

print("After OverSampling, counts of label '1': {}".format(sum(y_res==1)))
print("After OverSampling, counts of label '0': {}".format(sum(y_res==0)))
```

```
Before OverSampling, counts of label '1': 249
Before OverSampling, counts of label '0': 4861

After OverSampling, the shape of train_X: (7788, 8)
After OverSampling, the shape of train_y: (7788,)

After OverSampling, counts of label '1': 3894
After OverSampling, counts of label '0': 3894
```

Рисунок 4 – Використання SMOTE для збалансування даних

Конфузійні матриці для побудованих моделей машинного навчання показано на рисунках 5-

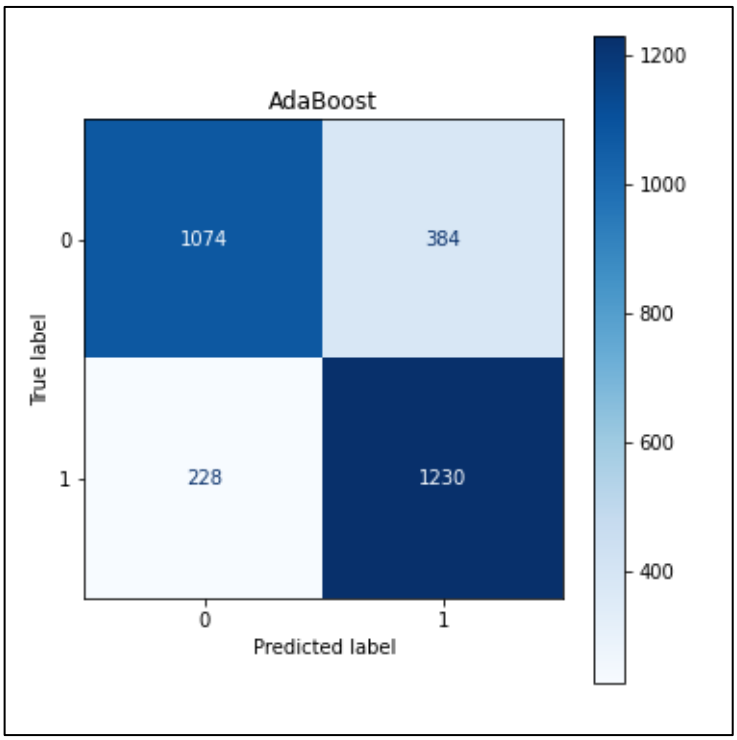


Рисунок 5 – Конфузійна матриця Ada Boost

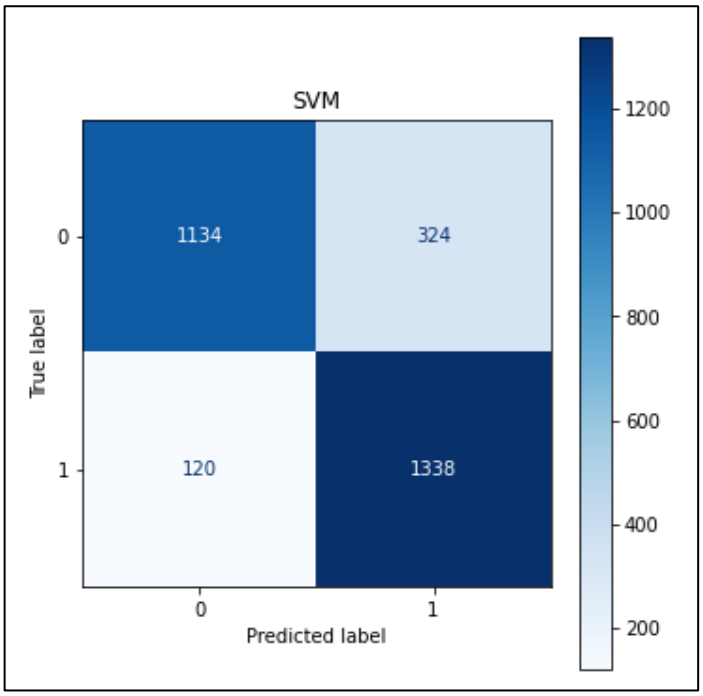


Рисунок 6 – Конфузійна матриця SVM

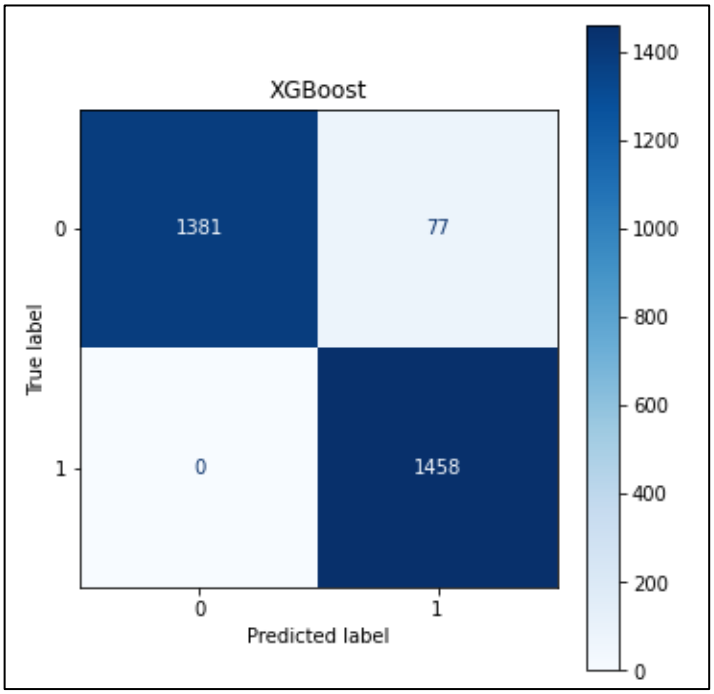


Рисунок 7 – Конфузійна матриця XG Boost

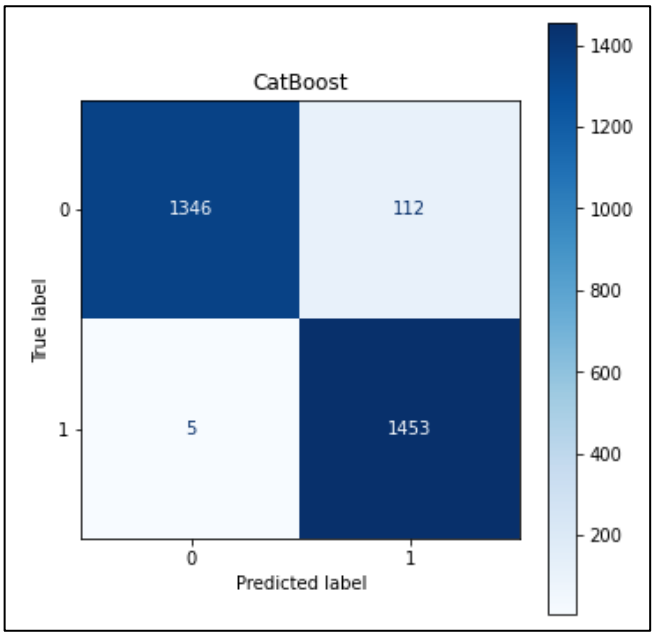


Рисунок 8 – Конфузійна матриця CatBoost

.

На рисунку 9 показано таблицю результатів моделей за метрикою accuracy_score.

| | Model | Training Accuracy | Testing Accuracy |
|---|---|---|---|
| 4 | XGBoost | 0.997648 | 0.973594 |
| 5 | CatBoost | 0.983833 | 0.959877 |
| 0 | KNN | 0.969871 | 0.944787 |
| 3 | Random Forest | 0.943416 | 0.926612 |
| 7 | KNearest | 0.943122 | 0.920439 |
| 2 | Decision Tree | 0.931511 | 0.919067 |
| 1 | SVM | 0.849500 | 0.831962 |
| 8 | AdaBoost | 0.797031 | 0.790123 |
| 6 | Logistic Regression | 0.785567 | 0.782236 |

Рис. 9 – Таблиця порівняння моделей

**Висновки**

Під час виконання роботи було розроблено інформаційну технологію для аналізу та передбачення стану хворих на інсульт з використанням різних моделей машинного навчання. Результати їх роботи були порівняні між собою і було визначено найбільш ефективну модель передбачення. У цілому, застосування різних моделей машинного навчання продемонструвало високий рівень точності в передбаченні хворих на інсульт.

У результаті побудови інформаційної технології передбачення хворих на інсульт з використанням моделей машинного навчання виявлено, що найвищу точність (0,974) за метрикою accuracy_score продемонструвала модель XGBoost.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Stroke, 2020 [Електронний ресурс] URL: https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113.
2. Fedesoriano Kaggle Stroke Prediction Dataset версія датасету – 2021 р.: [Електронний ресурс]. URL: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
3. SMOTE for Imbalanced Classification with Python 2021 [Електронний ресурс]. URL: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

**Козачко Олексій Миколайович** – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: lekoz80@gmail.com;

**Гонтковський Євгеній Юрійович** – студент групи 2ІСТ-22м, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: evgenhontkovskyi@gmail.com;

**Kozachko Oleksii M.** – Ph.D., associate professor of the System Analysis and Information Technologies Department, Vinnytsia National Technical University, Vinnytsia, e-mail: lekoz80@gmail.com

**Hontkovskyi Yevhenii Y. -** student of 2IST-22m group, Faculty of Intellectual Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: evgenhontkovskyi@gmail.com;