УДК 004.9:616.36

І. П. Гуцу
С. О. Жуков

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ СТАНУ ХВОРИХ НА ГЕПАТИТ

Вінницький національний технічний університет

Анотація

В роботі розроблено технологію аналізу та прогнозування стану хворих на гепатит. Для виконання поставленої задачі було використано такі моделі машинного навчання: виконання поставленої задачі є Logistic Regression, Random Forest Classifier, Gradient Boosting, SVM, Decision Tree Classifier.

Ключові слова: Python, гепатит, розвідувальний аналіз, захворювання

Abstract

In the work, the technology of analysis and forecasting of the condition of hepatitis patients was developed. The following machine learning models were used to perform the task: Logistic Regression, Random Forest Classifier, Gradient Boosting, SVM, Decision Tree Classifier.

Key words: Python, hepatitis, intelligence analysis, disease

Вступ

Щодня інформаційні технології стрімко розвиваються, призводячи до експоненційного зростання обсягу даних у світі інформаційних мереж. Це безпосередньо сприяє виникненню можливостей використання даних для різних цілей, зокрема для аналізу, класифікації та прогнозування. Завдяки цим даним можна проводити системний аналіз, виявляючи тенденції та ідентифікуючи фактори, які можуть оптимізувати функціонування різних систем.

Однією з актуальних проблем на сучасному етапі є гепатит, і вирішення цієї проблеми передбачає вдосконалення методів та засобів її виявлення. Раннє виявлення гепатиту має велике значення, оскільки це надає людям більше часу для вживання ефективних заходів [1-2].

Таким чином, використання інформаційних технологій для аналізу та обробки даних стає ключовим елементом у вдосконаленні підходів до діагностики та управління захворюванням. Інформаційна технологія аналізу дозволяє виявити не лише поточний стан справ, але і прогнозувати майбутні тенденції, що створює можливість вчасного реагування та оптимізації лікувальних процесів.

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.

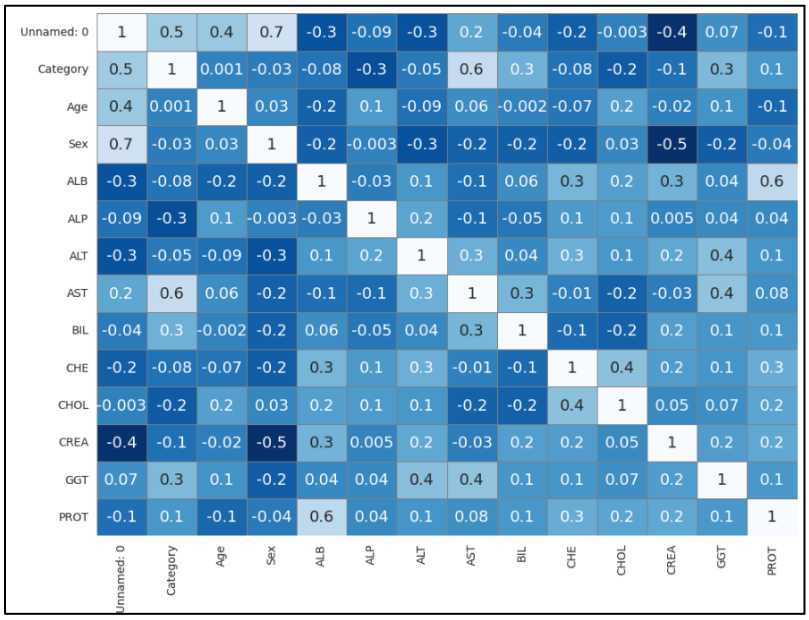

Рис. 2 – Кореляційна матриця датасету

На рисунку 3 показано пошук та видалення аномальних даних у датасеті для покращення подальших результатів аналізу.

[6]:

```
from scipy.stats import zscore
# Визначення порогу для Z-оцінки (зазвичай поріг 3 вважається значущим)
z_threshold = 3

# Обчислення Z-оцінок для кожної колонки
z_scores = zscore(df.select_dtypes(include=['float64']))

# Виявлення аномалій
anomalies = (z_scores > z_threshold).any(axis=1)

# Виведення кількості та видалення аномалій
print("Кількість аномалій:", anomalies.sum())
df = df[~anomalies].reset_index(drop=True)
```

```
Кількість аномалій: 53
```

Рис. 3 – Пошук та видалення аномальних значень

Матриці плутанини для побудованих моделей машинного навчання показано на рисунках 4-8.

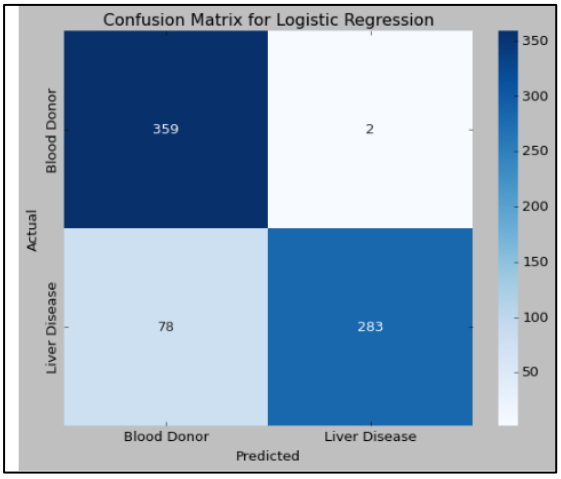

Рис. 4 – Матриця плутанини для моделі Logistic Regression

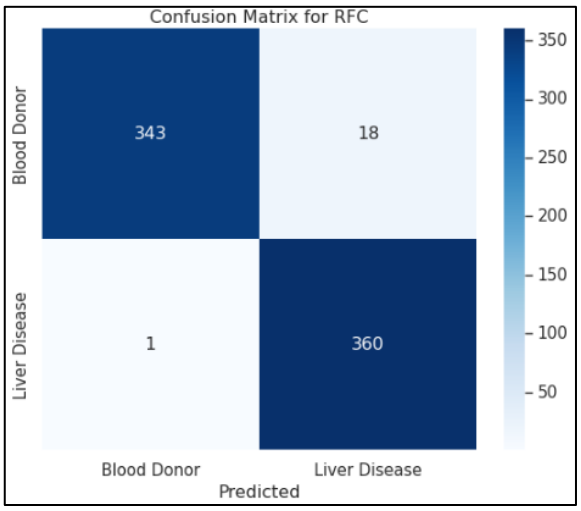

Рис. 5 – Матриця плутанини для моделі Random Forest Classifier

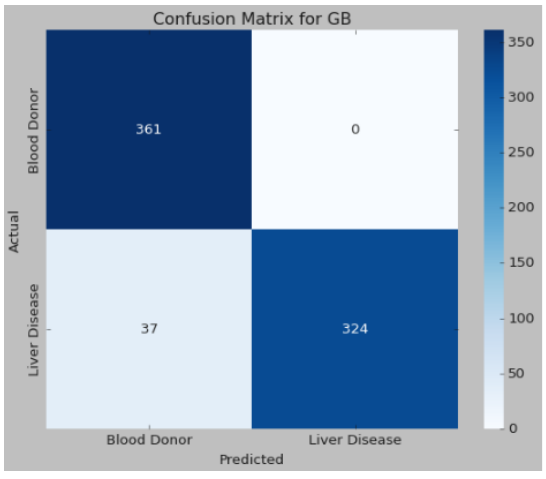

Рис. 6 – Матриця плутанини для моделі Gradient Boosting Classifier

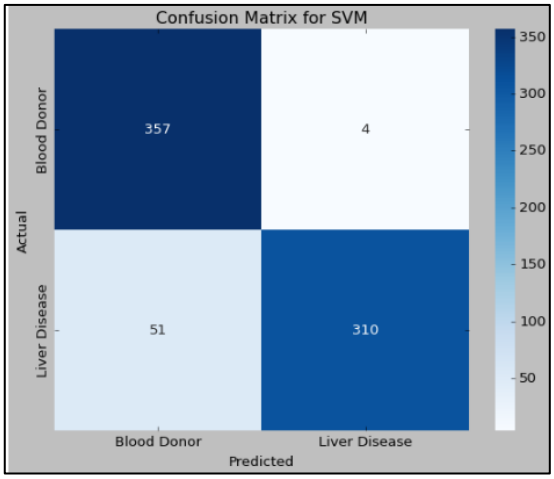

Рис. 7 – Матриця плутанини для моделі SVM

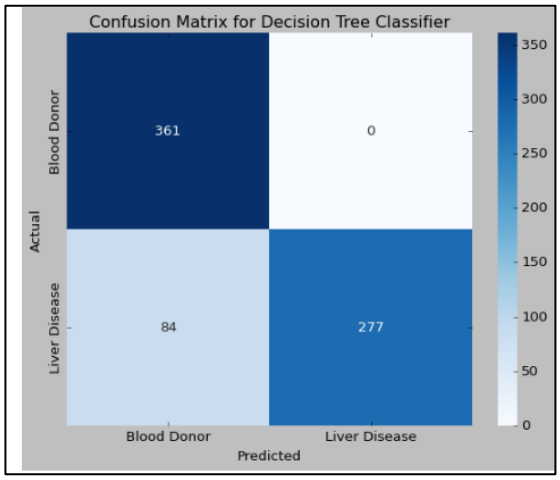

Рис. 8 – Матриця плутанини для моделі DecisionTreeClassifier

На рисунку 9 показано таблицю результатів моделей за метрикою r2_score.

	Модель	Точність на тренувальних даних	Точність на тестових даних
4	Decision Tree Classifier	0.989822	0.883657
0	Логістична регресія	0.974555	0.889197
3	Support Vector Machine Classifier	0.977099	0.923823
2	Gradient Boosting Classifier	0.918575	0.952941
1	Random Forest Classifier	0.954198	0.973684

Рис. 9 – Таблиця порівняння моделей

Висновки

Під час виконання роботи було реалізовано інформаційну технологію для аналізу та передбачення стану хворих на гепатит 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 стану хворих на гепатит.

У результаті побудови інформаційної технології для аналізу та передбачення стану хворих на гепатит з використанням моделей машинного навчання виявлено, що Random Forest Classifier продемонстрував найвищу точність прогнозування 0,973 за метрикою r2_score.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Кравченко, С. О., Кравченко, Т. О., Скрипник, Л. М. (2022). Гепатит: сучасні підходи до діагностики та лікування. Київ: Лань, 3-12.
2. Zhang, Y., Li, Y., Wang, M. (2022). A Deep Learning Model for Predicting the Outcomes of Hepatitis C Patients. Journal of Medical Internet Research.
3. Fedesoriano Kaggle Hepatitis C Prediction Dataset. версія датасету 2021 р. URL: https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset

Гуцу Ігор Петрович – студент групи 2ІСТ-22м, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: frewide.danmer@gmail.com

Жуков Сергій Олександрович– к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: sazhukov@gmail.com

Hutsu Ihor P. - student of 2IST-22m group, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: frewide.danmer@gmail.com

Zhukov Serhii O. - Ph.D., Assistant Professor of the Department of Systems Analysis and Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@gmail.com