

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ СТАНУ ХВОРИХ НА ГЕПАТИТ

¹Вінницький національний технічний університет

Анотація

У роботі представлено інформаційну технологію аналізу та передбачення стану хворих на гепатит із використанням методів машинного навчання. Система дозволяє здійснювати попередній аналіз медичних показників, виявляти закономірності та прогнозувати стан пацієнтів на основі біохімічних даних крові. Для моделювання застосовано алгоритми класифікації Random Forest, Support Vector Machine, Logistic Regression та Gradient Boosting, що забезпечують високу точність результатів. Розроблена технологія може бути використана для підтримки прийняття рішень у медичній діагностиці.

Ключові слова: машинне навчання, гепатит, класифікація, Python, аналіз даних, прогнозування стану пацієнтів.

Abstracts.

The paper presents an information technology for the analysis and prediction of the condition of patients with hepatitis using machine learning methods. The system enables preliminary analysis of medical indicators, detection of patterns, and prediction of patient conditions based on biochemical blood data. Classification algorithms such as Random Forest, Support Vector Machine, Logistic Regression, and Gradient Boosting were applied, providing high prediction accuracy. The developed technology can be used as a decision support tool in medical diagnostics.

Keywords: machine learning, hepatitis, classification, Python, data analysis, patient condition prediction.

Вступ

Гепатит є одним із найпоширеніших захворювань печінки, що має серйозні наслідки для здоров'я людини та потребує ранньої діагностики. Традиційні методи діагностики вимагають значних ресурсів і часу, тому актуальним є впровадження інтелектуальних систем, здатних аналізувати великі обсяги медичних даних та автоматично визначати ризики розвитку хвороби.

Метою роботи є розроблення інформаційної технології аналізу та передбачення стану хворих на гепатит на основі алгоритмів машинного навчання, що дозволить підвищити точність і швидкість діагностики.

Результати дослідження

Для реалізації поставленої мети використано відкритий набір даних Hepatitis C Prediction Dataset з платформи Kaggle [1], що містить демографічні та лабораторні показники пацієнтів. Попередня обробка даних включала очищення, нормалізацію, заповнення пропусків та усунення викидів за допомогою методу RobustScaler.

На наступному етапі було здійснено завантаження набору даних, після чого відобразили його частину для попереднього ознайомлення та перевірки коректності. Крім того, було отримано загальну інформацію про структуру датасету, що дозволило краще зрозуміти його склад та характерні особливості (див. рис. 1).

In [4]:

df

Out[4]:

	Unnamed: 0	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
610	611	3=Cirrhosis	62	f	32.0	416.6	5.9	110.3	50.0	5.57	6.30	55.7	650.9	68.5
611	612	3=Cirrhosis	64	f	24.0	102.8	2.9	44.4	20.0	1.54	3.02	63.0	35.9	71.3
612	613	3=Cirrhosis	64	f	29.0	87.3	3.5	99.0	48.0	1.66	3.63	66.7	64.2	82.0
613	614	3=Cirrhosis	46	f	33.0	NaN	39.0	62.0	20.0	3.56	4.20	52.0	50.0	71.0
614	615	3=Cirrhosis	59	f	36.0	NaN	100.0	80.0	12.0	9.07	5.30	67.0	34.0	68.0

615 rows x 14 columns

Рисунок 1 – Діаграма класів інформаційної веб-системи для автоматизації особистого тайм-менеджменту.

Датасет містить лабораторні показники крові донорів та пацієнтів з гепатитом С, а також демографічні дані, такі як вік і стать. Джерелом даних є UCI Machine Learning Repository.

Усі атрибути, крім категорії (Category) та статі (Sex), є числовими. Перші чотири атрибути описують основну інформацію про пацієнта: унікальний ідентифікатор (ID), діагноз, вік у роках і стать (чоловіча або жіноча).

Наступні десять атрибутів — це лабораторні показники крові, які включають рівні таких речовин, як альбумін (ALB), лужна фосфатаза (ALP), аланінова та аспартатамінотрансферази (ALT, AST), білірубін (BIL), холінестераза (CHE), холестерин (CHOL), креатинін (CREA), гамма-глутамілтрансфераза (GGT) та загальний білок (PROT).

Цільовою змінною для задачі класифікації є атрибут Category, який розподіляє пацієнтів на кілька груп: здорові донори крові, підозрювані донори та різні стадії хвороби — гепатит, фіброз і цирроз.

Здійснено розвідувальний аналіз, у тому числі побудовано матрицю кореляції показників (рис. 2). Аналіз сирих даних виявив закономірності, зокрема значний вплив параметра AST на виникнення гепатиту, що дозволило ідентифікувати ключові фактори ризику та закласти основу для моделювання.

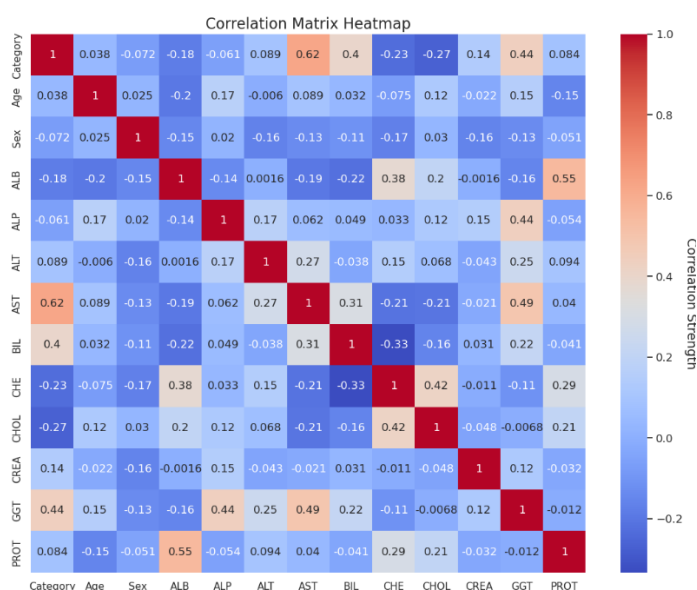


Рисунок 2 – Матриця кореляції показників

На основі проведеного аналізу побудовано кілька моделей класифікації: Random Forest, Support Vector Machine, Logistic Regression та Gradient Boosting [2-5]. Порівняльний аналіз показників точності дозволив визначити оптимальну модель для практичного застосування.

Побудовано моделі класифікації (рис. 3) з використанням асамблевого підходу, який підвищив точність до 0.95. Найкращий результат продемонстрував Random Forest, що домінував у голосуванні асамблевого моделювання. Gradient Boosting також показав високу ефективність (93.50% на тренуванні та 94% на тесті), підтвердивши надійність ансамблевих методів.

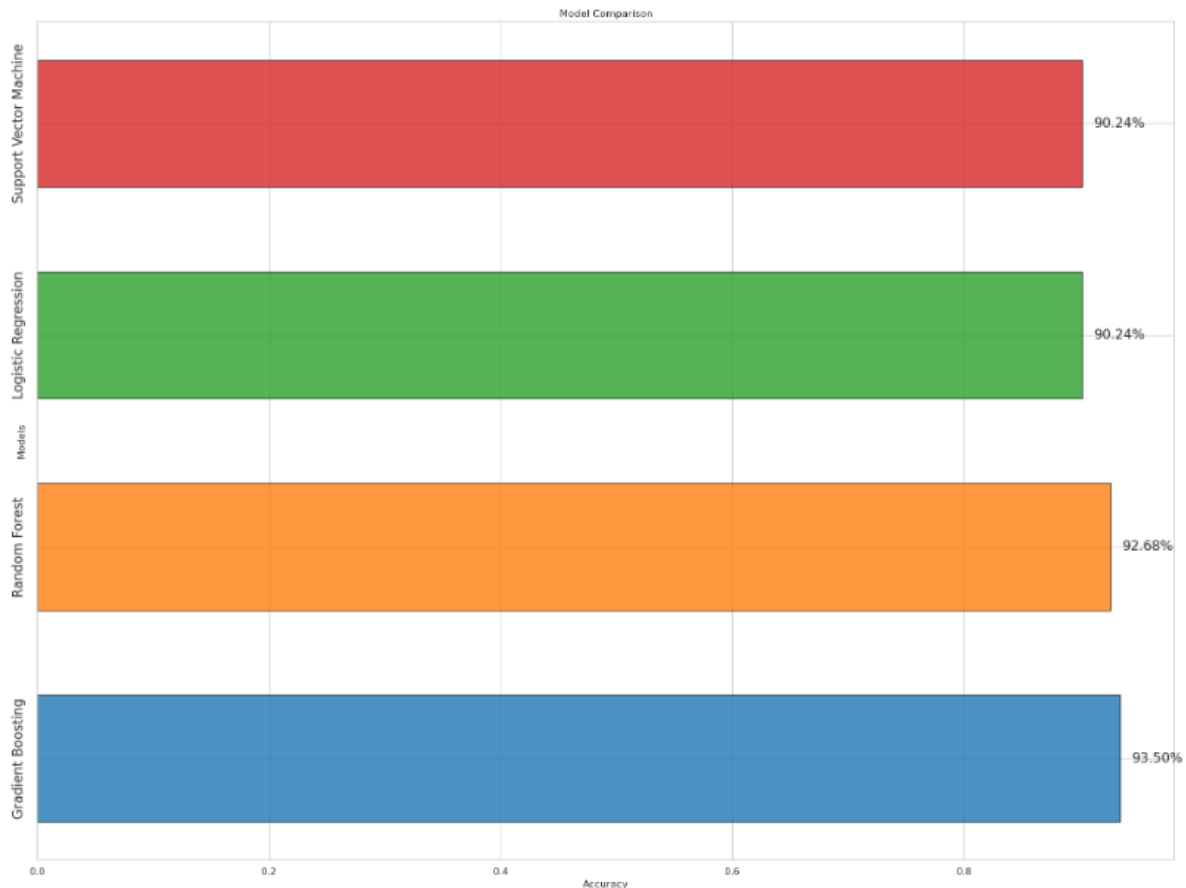


Рисунок 3 – Результати роботи моделей

Розроблена інформаційна технологія забезпечує ефективну обробку медичних даних, виявлення прихованих закономірностей і надання прогнозів щодо ймовірності розвитку хвороби.

Висновки

Розроблена інформаційна технологія аналізу та передбачення стану хворих на гепатит показала високу ефективність застосування алгоритмів машинного навчання для класифікації медичних даних. Система може бути інтегрована в медичні інформаційні комплекси для підтримки лікарів у процесі діагностики та моніторингу пацієнтів. Подальші дослідження передбачають розширення набору даних і використання нейронних мереж для підвищення точності прогнозування.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Kaggle. Hepatitis C Prediction Dataset [Електронний ресурс]. — Режим доступу: <https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset>.
2. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. — Journal of Machine Learning Research, 2011.

3. McKinney W. Data Analysis with Python and pandas. — O'Reilly Media, 2022.
4. Chollet F. Deep Learning with Python. — Manning, 2021.
5. Наука про дані: машинне навчання та інтелектуальний аналіз даних – Електронний навчальний посібник / В. Б. Мокін, М. В. Дратований – Вінниця : ВНТУ, 2024. – 263 с.

Поліщук Денис Миколайович – студент групи 2ІСТ-24м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця/

Євгеній Миколайович Крижановський – канд. техн. наук, доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: kruzhan@gmail.com

Ігор Миколайович Штельмах – канд. техн. наук, асистент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: igor.shtelmakh@vntu.edu.ua.

Polishchuk Denys M. - student of Faculty of Intelligent Information Technology and Automation, 2IST-24m, Vinnytsia National Technical University, Vinnytsia/

Kryzhanovsky, Evgeniy M. - candidate of technical sciences, associate professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: kruzhan@gmail.com

Shtelmah, Igor M. – candidate of technical sciences, assistant professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: igor.shtelmakh@vntu.edu.ua.