УДК 004.6

С. Д. Неволя
С. О. Жуков

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ РАКУ ЛЕГЕНІВ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання для інформаційної технології передбачення раку легень методами машинного навчання. Було проведено аналіз датасету та його ознак.

Ключові слова: рак легенів, інформаційні технології, машинне навчання, аналіз даних, передбачення, ознаки, передбачення раку легенів.

Abstract

The work is devoted to the preparation and intelligence analysis of data for further use for information technology for the prediction of lung cancer using machine learning methods. An analysis of the dataset and its features was carried out.

Keywords: Lung Cancer, Information Technology, Machine Learning, Data Analysis, Predictions, Signs, Lung Cancer Predictions.

Вступ

Кожного дня технології у світі стрімко розвиваються, те що умовно кажучи учора було фантастикою, завтра уже реальність. Особливо це відчувається у сфері інформаційних технологій, які стали основою сьогодення. З їхньою допомогою відкриваються нові можливості для вирішення завдань, які раніше вважались неможливими. Одним із таких завдань, де інформаційні технології допомагають вирішувати проблеми та завдання, є медицина. Наприклад для встановлення діагнозу при наявності певних ознак та аналізів чи прогнозування вірогідності захворіти.

Одним із таких захворювань, які можливо спрогнозувати, є рак легенів. Передбачення даного захворювання дасть змогу завчасно підготуватись до такого розвитку подій, зменшити ризики такого захворювання та в цілому вжити найбільш ефективних заходів для запобігання захворювання.

Виходячи з цього, використання інформаційних технологій для обробки та аналізу даних дає можливості для удосконалення діагностичних методів, методів передбачення та запобігання.

Розвідувальний аналіз

Для проведення аналізу було обрано набір даних, що має назву «Lung Cancer Prediction» та опублікований користувачем «The Devastator» та має відкритий доступ для загального використання на платформі Kaggle [1]. Даний датасет містить у собі широкий спектр даних про пацієнтів лікарень з раком легенів, які надає журнал Nature Medicine. Серед стовпців із параметрами можна виділити такі як вік, вплив забрудненого повітря, куріння, вживання алкоголю, професійні ризики, генетичні ризики та ще 19 інших категорій (рис. 1).

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue
0	0	P1	33	1	2	4	5	4	3	2	...	3
1	1	P10	17	1	3	1	5	3	4	2	...	1
2	2	P100	35	1	4	5	6	5	5	4	...	8
3	3	P1000	37	1	7	7	7	7	6	7	...	4
4	4	P101	46	1	6	8	7	7	7	6	...	3
...	...	...	...	...	...	...	...	...	...	...	...	...
995	995	P995	44	1	6	7	7	7	7	6	...	5
996	996	P996	37	2	6	8	7	7	7	6	...	9
997	997	P997	25	2	4	5	6	5	5	4	...	8
998	998	P998	18	2	6	8	7	7	7	6	...	3
999	999	P999	47	1	6	5	6	5	5	4	...	8

Рис. 1. Приклад ознак пацієнтів, що містить набір даних

Даний датасет включає у себе такі колонки:

- Age: Вік пацієнта. (Numeric)
- Gender: Стать пацієнта. (Categorical)
- Air Pollution: рівень впливу забруднення повітря на пацієнта. (Categorical)
- Alcohol use: рівень вживання алкоголю пацієнтом. (Categorical)
- Dust Allergy: рівень алергії на пил у пацієнта. (Categorical)
- OccuPational Hazards: Рівень професійних ризиків пацієнта. (Categorical)
- Genetic Risk: рівень генетичного ризику пацієнта. (Categorical)
- chronic Lung Disease: рівень хронічного захворювання легенів у пацієнта. (Categorical)
- Balanced Diet: рівень збалансованості дієти пацієнта. (Categorical)
- Obesity: рівень ожиріння пацієнта. (Categorical)
- Smoking: рівень куріння пацієнта. (Categorical)
- Passive Smoker: рівень пасивного куріння пацієнта. (Categorical)
- Chest Pain: рівень болю в грудях пацієнта. (Categorical)
- Coughing of Blood: Рівень відкашлювання крові пацієнта. (Categorical)
- Fatigue: рівень втоми пацієнта. (Categorical)
- Weight Loss: рівень втрати ваги пацієнта. (Categorical)
- Shortness of Breath: рівень задишки пацієнта. (Categorical)
- Wheezing: Рівень хрипів у пацієнта. (Categorical)
- Swallowing Difficulty: рівень труднощів ковтання пацієнта. (Categorical)
- Clubbing of Finger Nails: Рівень збивання нігтів пацієнта. (Categorical)

Проведено попереднє очищення даних. Підготувавши дані створимо для них матрицю кореляції, що дасть змогу виявити явну залежність та зв’язки між різними параметрами та ознаками.

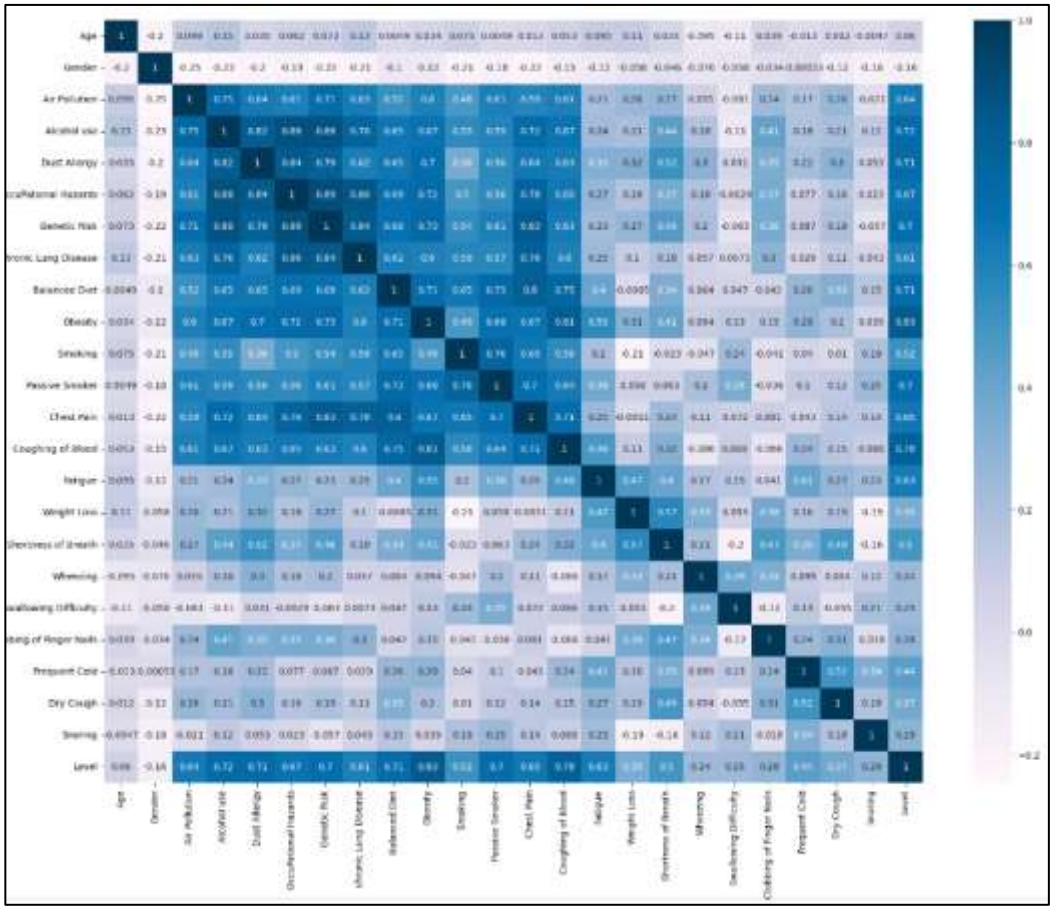

Рис. 2. Матриця кореляції

З рисунку 2 можна зробити висновок що найбільше рівень ризику захворіти раком легенів залежить від таких параметрів як ожиріння (Obesity), рівень вживання алкоголю (Alcohol use), генетичні ризики (Genetic Risk), рівень відкашлювання крові (Coughing of Blood).

Далі візуально відобразимо кількість наявних даних, відповідно до окремих параметрів (рис. 3).

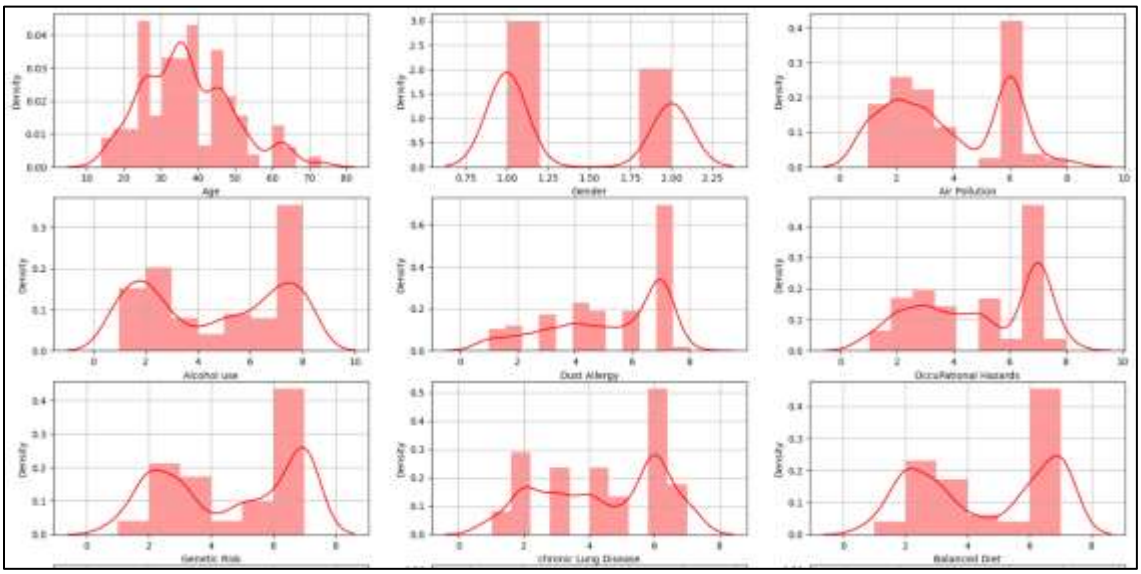

Рис. 3. Візуалізація кількості даних у вибірці по ознаках

З рисунку 3 можемо зробити висновок що вибірка по віковій категорії більше припадає на середній вік пацієнтів, від 25 до 45 років, вибірка по статевій ознаці припадає більшістю на чоловіків (позначена

у датасеті як 1), і переважна більшість даних із високими впливами забрудненого повітря, вживання алкоголю, алергією на пил, генетичними ризиками та хронічними захворюваннями легень.

Далі візуалізуємо кількість пацієнтів у відповідності між рівнем ризику (де 0 це низький рівень, 1 середній рівень, 2 високий рівень ризику), та між віком (рис. 4) й статевою ознакою (рис. 5).

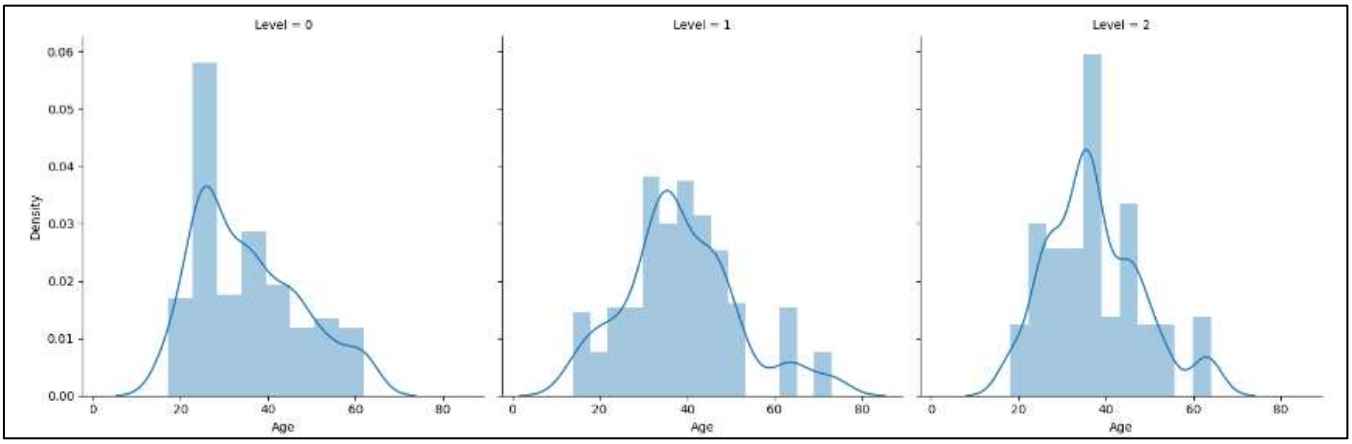

Рис. 4. Візуалізація ризику захворіти раком легенів відповідно до віку

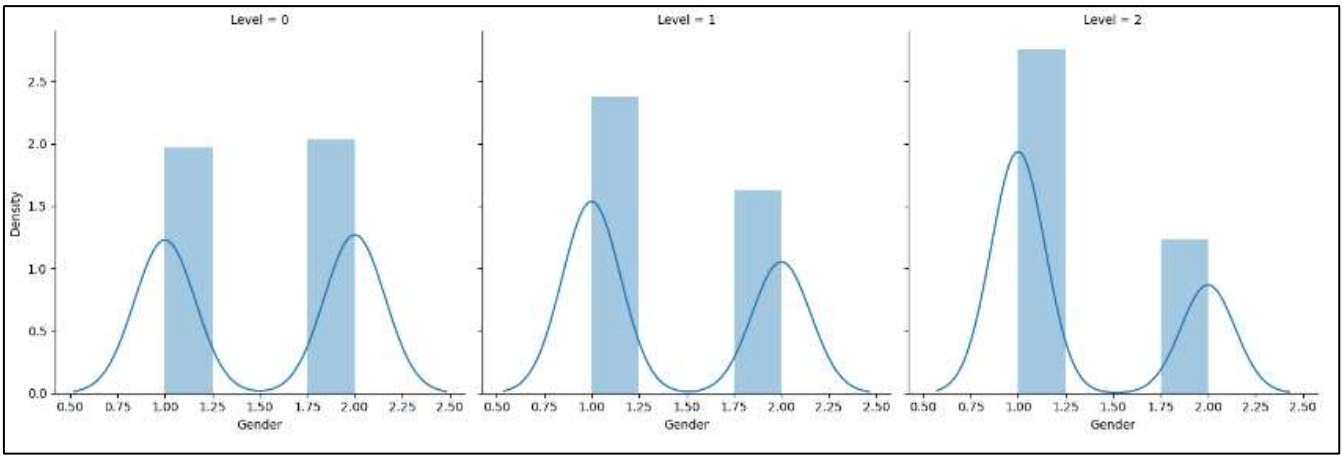

Рис. 5. Візуалізація ризику захворіти раком легенів відповідно до гендерної ознаки

З візуалізації видно, що у нашій вибірці кількість людей віком 20 років має найменші ризики захворіти, середні ризики рівномірно розподілені між людьми віком між від 30 до 45 років, тоді як кількість людей із високими ризиками захворювання припадає більшістю на пацієнтів віком 35-40 років. Щодо гендерної ознаки можемо бачити, що низький ризик майже рівномірно розподілений між жінками та чоловіками із невеликим відхиленням у сторону жінок, тоді як у відповідності до середнього та високого ризиків чоловіки мають явну перевагу.

Виконавши візуалізацію кількості даних відповідно до ризиків захворіти за допомогою кругової діаграми бачимо що дані за цим параметром розподілені рівномірно (рис. 6).

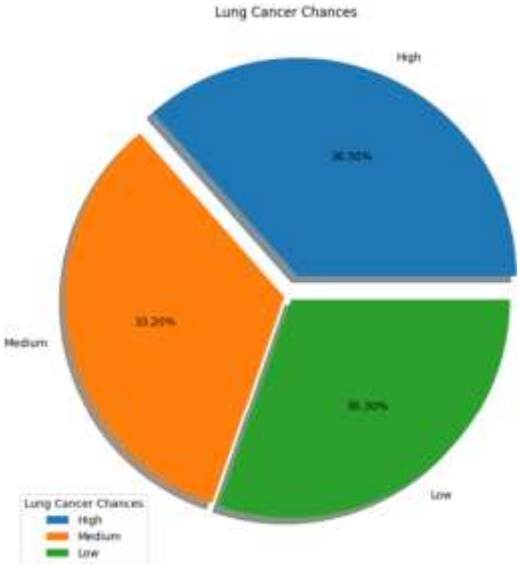

Рис. 6. Кругова діаграма ризиків захворюваності

Висновки

При розвідувальному аналізі набору даних «Lung Cancer Prediction», що містить у собі інформацію про параметри та ризики захворювання на рак легенів було досліджено вплив різник ознак на показник ризику. Побудовано матрицю кореляції, яка показує залежність між усіма параметрами та зроблено висновки щодо залежності параметру рівня ризику (Level) до інших ознак.

Далі було побудовано стовпчасті діаграми для візуалізації кількості даних відповідно до окремих ознак, а саме вік, стать, рівень впливу забрудненого повітря, рівень вживання алкоголю, рівень алергії на пил і т. ін.

Також було досліджено відповідності між рівнем ризику, та між віком й статевою ознакою та виконано візуалізацію кількості даних відповідно до ризиків захворіти за допомогою кругової діаграми, що показало нам рівномірне розподілення даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Lung Cancer Prediction Dataset. Kaggle. 2023 [Електронний ресурс]. – Режим доступу: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link
2. Pandas Getting started. 2024 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
3. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
4. Seaborn Tutorial. 2023 [Електронний ресурс]. – Режим доступу: https://seaborn.pydata.org/tutorial.html

Неволя Сергій Дмитрович – студент групи 2ІСТ-20б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: nevolya2003@gmail.com

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

Nevolya Serhii Dmytrovych. - student of Faculty of Intelligent Information Technology and Automation, 2IST-20b, Vinnytsia National Technical University, Vinnytsia, e-mail nevolya2003@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