УДК 004.6

С. Д. Неволя
С. О. Жуков

ТРЕНУВАННЯ МОДЕЛЕЙ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ РАКУ ЛЕГЕНЬ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та тренуванню моделей машинного навчання для подальшого використання для інформаційної технології передбачення раку легень методами машинного навчання. Була проведена обробка датасету та його ознак, що включало в себе генерування нових ознак та натреновано нові моделі.

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

Abstract

The work is devoted to the preparation and training of machine learning models for further use for information technology for lung cancer prediction by machine learning methods. The dataset and its features were processed, which included the generation of new features and the training of new models.

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

Вступ

Технології у світі кожного дня 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. Завчасне передбачення даної хвороби надає пацієнтам змогу збільшити свої шанси на своєчасне лікування, адже головною причиною високої смертності від подібних хвороб є пізня діагностика, коли лікарі і сучасна медицина не в силах нічого вдіяти.

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 генерацію нових ознак за допомогою уже наявних, для цього використовується функція PolynomialFeatures та за основу беруться основні ознаки, які корелюються між собою, а саме Air Pollution, Alcohol use, OccuPational Hazards та Genetic Risk.

Passive_Smoker Genetic_Risk	Passive_Smoker Obesity	Passive_Smoker Alcohol_use	Passive_Smoker OccuPational_Hazards	Genetic_Risk Obesity	Genetic_Risk Alcohol_use	Genetic_Risk OccuPational_Haz
6.0	8.0	8.0	8.0	12.0	12.0	12.0
15.0	21.0	15.0	15.0	35.0	25.0	25.0
42.0	49.0	49.0	49.0	42.0	42.0	42.0
49.0	49.0	56.0	49.0	49.0	56.0	49.0
15.0	21.0	15.0	15.0	35.0	25.0	25.0
...	...	...	...	...	...	...
56.0	56.0	56.0	56.0	49.0	49.0	49.0
56.0	56.0	64.0	56.0	49.0	56.0	49.0
15.0	21.0	15.0	15.0	35.0	25.0	25.0
49.0	49.0	56.0	49.0	49.0	56.0	49.0
15.0	21.0	15.0	15.0	35.0	25.0	25.0

Рис. 2. Згенеровані нові ознаки

З рисунку 2 можна зробити висновок що нові ознаки згенерувались коректно та дали нам прийнятний результат.

Далі візуально наведемо приклад як відбувається навчання однієї з моделей (рис. 3).

```
param_svm = {
    'C': [0.1],              # маленькі C = сильніша регуляризація
    'kernel': ['rbf'],
    'gamma': ['scale', 'auto', 0.001, 0.01, 0.1, 1]              # авто
ння γ
}
```

```
%%time
model_tuning = GridSearchCV(estimator=SVC(), param_grid=param_svm, cv=5)
model_tuning.fit(x_train, y_train)

best_params = model_tuning.best_params_
print("Best Parameters:", best_params)
model_svm = SVC(**best_params)
model_svm.fit(x_train, y_train)
```

Рис. 3. Візуалізація кількості даних у вибірці по ознаках

З рисунку 3 можемо бачити що для конкретно моделі SVM використовується функція GridSearchCV, яка допомагає обрати найкращу комбінацію параметрів, яка дає найвищий результат точності під час навчання моделі. Принцип роботи даної функції полягає в перебиранні усіх можливих комбінацій заданих параметрів, це дозволяє знайти найкращу, але в той же час збільшення кількості заданих параметрів суттєво збільшує кількість комбінацій що впливає на швидкість роботи програми, тому при роботі даною функцією рекомендується використовувати помірну кількість параметрів для оптимізації роботи усієї програми.

Далі після отримання найкращої комбінації параметрів ми перевіряємо точність моделі на тестових даних взятих із датасету (рис. 4). Також для порівняння точності моделі на тренувальних даних та тестових даних програма виводить відповідні результати (рис. 5).

```
# Train
train_predictions = model_svm.predict(x_train)
r2_train = r2_score(y_train, train_predictions)
f1_train = f1_score(y_train, train_predictions, average = 'micro')

# Test
test_predictions = model_svm.predict(x_test)
r2_test = r2_score(y_test, test_predictions)
f1_test = f1_score(y_test, test_predictions, average = 'micro')

#Result
performTrain(train_predictions)
performTest(test_predictions)
```

Рис. 4. Перевірка моделі на тестових даних

```
Train Data Metrics:
Precision :  0.9985754985754985
Recall :  0.9985754985754985
Accuracy :  0.9985754985754985
F1 Score :  0.9985754985754985
R2 Score :  0.9978762796776302


Test Data Metrics:
Precision :  0.9848484848484849
Recall :  0.9848484848484849
Accuracy :  0.9848484848484849
F1 Score :  0.9848484848484849
R2 Score :  0.9358923624851603
```

Рис. 5. Результати навчання моделі на тренувальних та на тестових даних

За результатами навчання моделі бачимо що результати вона показала чудові, оскільки наша задача це задача класифікації для оцінки результативності моделей в першу чергу доцільніше звертати увагу на показник метрики F1 Score, за ним модель на тренувальних даних показала точність 0.998 а на тестових 0.984. Як бачимо різниця між точністю на тренувальних даних та на тестових є мінімальною, що означає що модель під час навчання не просто запам’ятовує дані, а може їх узагальнювати на раніше невідомі дані, що є гарним показником якості такої моделі.

	model	f1_train	f1_test	r2_train	r2_test	short
0	RandomForest Classifier	0.992877	0.974747	0.989381	0.964385	RF
1	ADABoost Classifier	0.990028	0.984848	0.985134	0.978631	ADA
2	ExtraTrees Classifier	0.924501	0.914141	0.887443	0.878908	ExT
3	DecisionTree Classifier	0.960114	0.934343	0.895938	0.843292	DT
4	XGB Classifier	0.964387	0.919192	0.902309	0.821923	XGB
5	MLP Classifier	0.990028	0.984848	0.985134	0.978631	MLP
6	LR Classifier	0.97151	0.949495	0.957526	0.928769	LR
7	SVM Classifier	0.998575	0.984848	0.997876	0.935892	SVM
8	KNN Classifier	0.995726	0.994949	0.993629	0.992877	KNN
9	LGBM Classifier	0.988604	0.984848	0.98301	0.978631	LGBM

Рис. 6. Результати тренування усіх моделей

За результатами усіх моделей зображеного на рисунку 6, можемо визначити що найкращою з них усіх є модель KNN із результатом точності 0.994 за метрикою F1 Score, попри це слід також відмітити що інші моделі також мають високі показники точності, що може говорити про якісні дані в датасеті.

Висновки

При обробці 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 та згенеровано нові ознаки.

Далі було побудовано десять моделей машинного навчання, було проведено тюнінг їх параметрів та отримано хороші результати навчання. З усіх моделей було обрано найкращу, а саме KNN, із точністю 0.994 за метрикою F1 Score.

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. 2025 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
3. Matplotlib Pyplot Documentation. 2025 [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-24м, 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-24m, 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