Святослав Кицара, Ігор Пірко*
[1] НЛТУ України, Львів, Україна, 23kytsara.s@nltu.lviv.ua
[2] НЛТУ України, Львів, Україна. ORCID: *0009-0008-2378-2929*,
pirko@nltu.edu.ua

Розроблення інформаційної системи медичного страхування пацієнтів

Анотація. В даній роботі використано мову програмування Python та ряд бібліотек, таких як Pandas, Seaborn та Scikit-Learn для аналізу та моделювання даних щодо медичного страхування. Робота містить в себе імпорт та обробку даних, візуалізацію розподілу ознак, побудову графіків впливу різних факторів на вартість страхування, а також створення та оцінку моделі лінійної регресії для передбачення медичних витрат. Проведено широкий аналіз залежності між різними ознаками та вартістю страхування, що дає змогу краще розуміти фактори, які впливають на ціни медичного страхування.

Ключові слова – медичне страхування, машинне навчання, Python, Pandas, Matplotlib, Seaborn, Scikit-Learn.

Зростання вартості медичного страхування та пошук ефективних методів його визначення стають актуальними завданнями у сучасному світі. Аналіз факторів, які впливають на вартість медичного страхування, є критичним для страхових компаній та інших учасників ринку для раціонального формування тарифів. Висвітлення залежності між різними ознаками та вартістю страхування дає змогу краще розуміти структуру цін та ідентифікувати ключові фактори впливу. Використання методів машинного навчання, зокрема моделей лінійної регресії, у визначенні вартості страхування відкриває нові можливості для точніших та обґрунтованих прогнозів. Підвищення точності визначення вартості страхування сприяє покращенню фінансової стійкості страхових компаній та забезпеченню економічної стабільності клієнтів [1].

Математична модель. Дане дослідження спрямоване на побудову прогнозу для пацієнта щодо того, чи звернеться він у найближчий час до поліклініки із захворюваннями. Ця модель прогнозування витрат на медичне страхування буде використовувати алгоритм лінійної регресії для прогнозування витрат на медичне страхування особи на основі наданих даних.

* науковий керівник

Бібліотека Pandas використовується для роботи з даними у вигляді таблиць, Seaborn надає інтерфейс для створення інформативних графіків на основі даних з Pandas. Бібліотека Matplotlib дає змогу створювати графіки та діаграми. Функція train_test_split з бібліотеки Scikit-Learn використовується для розділення набору даних на тренувальний і тестовий, а клас LinearRegression з цієї ж бібліотеки – для побудови моделі лінійної регресії, що використовується для моделювання лінійних залежностей між змінними [2].

Таблиця1. Датасет моделі медичного страхування

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Для побудови діаграм використано бібліотеки Matplotlib та Seaborn, які відображають різні розподіли. На рис. 1 представлено графіки, кожен з яких візуалізує зв’язок між категоріальними ознаками та числовою ознакою (витрати на страхування). Вони дають змогу побачити розподіл витрат на страхування для різних значень категоріальних ознак. З використанням графіків можна порівняти вплив різних категоріальних ознак на числову ознаку [3].

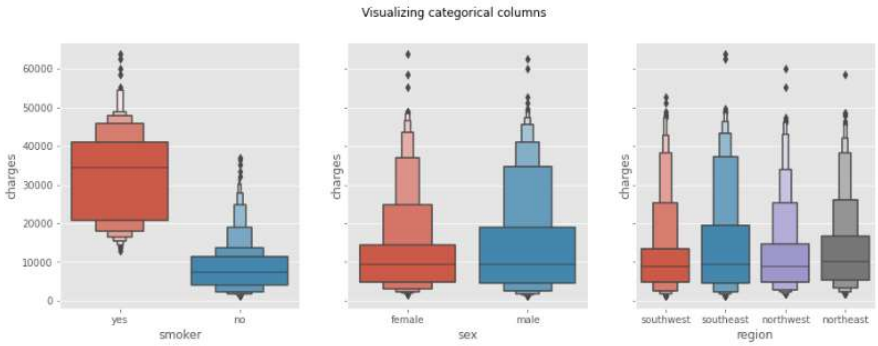

Рисунок 1. Візуалізація зв’язку між категоріальними ознаками та витратами на страхування

З допомогою теплової карти візуалізують коефіцієнти кореляції між числовими стовпцями датафрейму (рис. 2). Це дає змогу зрозуміти взаємозв’язки між різними ознаками у наборі даних.

* науковий керівник

Рисунок 2. Теплова карта кореляції ознак

Об’єкт лінійної регресії буде використовуватися для навчання моделі на навчальному наборі та для здійснення передбачень на тестовому наборі. Знайдене лінійне відношення потім може бути використане для здійснення передбачень на нових даних. Після створення об’єкта LinearRegression використовують методи цього об’єкта для навчання моделі та отримання передбачень.

Програмна реалізація. Створюють новий датафрейм для нового клієнта з вказаними ознаками (вік, індекс маси тіла, кількість дітей, чи є курцем, регіон). Для цього створюють словник, де ключі – це назви ознак, а значення – конкретні значення ознак для нового клієнта. Кожний ключ у словнику стає назвою стовпця, а відповідне значення – значенням в цьому стовпці для нового клієнта. Виводять отриманий датафрейм, який представляє нового клієнта з вказаними характеристиками. Тепер його можна використовувати для здійснення передбачень за допомогою раніше навченої моделі лінійної регресії.

Таблиця 2. Вхідні ознаки нового клієнта

	age	bmi	children	smoker	region
0	50	25	2	1	2

Далі використовують навчену модель лінійної регресії, щоб здійснити передбачення страхових витрат для нового клієнта. Для цього застосовують метод predict до моделі лінійної регресії для передбачення вартості медичного страхування для нового клієнта, якого представлено в датафреймі. Виводять передбачену вартість медичного страхування для нового клієнта на екран. Це дають змогу отримати приблизне значення вартості медичного страхування для ново-

* науковий керівник

го клієнта на основі навченої моделі лінійної регресії та введених характеристик для цього клієнта. У контексті медичного страхування це число може вказувати на річні витрати клієнта на медичні послуги, які можуть бути покриті страховкою. Отримане значення вказує на приблизну вартість медичного страхування для нового клієнта, яка розрахована за допомогою навченої моделі лінійної регресії та введених характеристик клієнта.

Висновок. Інформаційну систему розроблено мовою програмування Python з використанням ряду бібліотек, таких як Pandas, Seaborn та Scikit-Learn для аналізу та моделювання даних щодо медичного страхування. Вона містить в себе імпорт та обробку даних, візуалізацію розподілу ознак, побудову графіків впливу різних факторів на вартість страхування, а також створення та оцінку моделі лінійної регресії для передбачення медичних витрат. Проведено аналіз залежності між різними ознаками та вартістю страхування, що дає змогу краще розуміти фактори, які впливають на ціни медичного страхування. Оцінено ефективність моделі за допомогою коефіцієнта детермінації R^2. Продемонстровано можливість використання навченої моделі для передбачення вартості медичного страхування для нового клієнта на основі введених характеристик.

Список використаних літературних джерел

[1] Hanafy, M., & Mahmoud, O.M. (2021). Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models. International Journal of Innovative Technology and Exploring Engineering, 10, 137–143.

[2] Bhardwaj, N., & Anand, R. (2020). Health Insurance Amount Prediction. International Journal of Engineering Research, 9, 1008–1011.

[3] Goundar, S., Prakash, S., Sadal, P., & Bhardwaj, A. (2020). Health Insurance Claim Prediction Using Artificial Neural Networks. International Journal of System Dynamics and Applications, 9, 40–57.

Development of a patient health insurance information system

Sviatoslav Kitsara, Igor Pirko*

* науковий керівник

This paper uses the Python programming language and a number of libraries such as Pandas, Seaborn, and Scikit-Learn to analyze and model health insurance data. The work includes importing and processing data, visualizing the distribution of features, plotting the impact of various factors on the cost of insurance, and creating and evaluating a linear regression model to predict medical costs. An extensive analysis of the relationship between various characteristics and the cost of insurance was conducted, which allows for a better understanding of the factors that affect health insurance prices.

* науковий керівник