

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ РІВНЯ ЩАСТЯ У КРАЇНАХ СВІТУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Дана робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання у інформаційній технології передбачення рівня щастя у країнах світу методами машинного навчання. Було проведено аналіз датасету World Happiness Report, що містить соціально-економічні характеристики країн, та досліджено залежності між ними і показником щастя.

Ключові слова: рівень щастя, інформаційні технології, машинне навчання, аналіз даних, передбачення, ознаки, передбачення рівня щастя у країнах світу.

Abstract

The paper is devoted to the preparation and exploratory data analysis for further use in the information technology of predicting the happiness level in countries worldwide using machine learning methods. The World Happiness Report dataset, which includes socio-economic indicators, was analyzed to discover relationships between these indicators and the happiness score.

Keywords: happiness level, information technology, machine learning, data analysis, prediction, features, prediction of happiness level in countries worldwide.

Вступ

У сучасному світі рівень добробуту населення та якість життя є пріоритетними аспектами соціального розвитку. Одним із важливих індикаторів цих аспектів є рівень щастя громадян, що все частіше стає предметом досліджень у глобальному масштабі. Звіт про щастя у світі (World Happiness Report) дозволяє комплексно оцінити цей показник, спираючись на соціально-економічні характеристики країн.

Зі зростанням обсягів доступних даних та розвитком технологій обробки інформації, з'являються нові можливості для аналізу складних соціальних явищ. Застосування інформаційних технологій, зокрема методів машинного навчання, створює умови для виявлення закономірностей, що визначають рівень щастя, та побудови моделей, які здатні передбачати його зміну залежно від різних факторів.

У цьому дослідженні проведено розвідувальний аналіз даних на основі відкритого набору World Happiness Report, доступного на платформі Kaggle [1]. Основною метою є виявлення ключових факторів, що впливають на щастя населення, та формування бази для створення ефективної інформаційної технології прогнозування.

Розвідувальний аналіз

Для проведення аналізу було обрано набір даних під назвою World Happiness Report, який має відкритий доступ для загального використання на платформі Kaggle [1]. Даний датасет містить інформацію про рівень щастя в різних країнах світу, а також низку соціально-економічних показників, що потенційно впливають на цей рівень. Загалом датасет охоплює понад 150 країн і містить 9 ключових ознак, що використовуються для розрахунку індексу щастя (рис. 1).

Основні атрибути датасету:

- Country – назва країни;
- Happiness Rank – рейтинг країни на основі оцінки щастя;
- Happiness Score – оцінка рівня щастя населення;
- GDP per capita – валовий внутрішній продукт на душу населення;
- Social Support – рівень соціальної підтримки;

- Healthy Life Expectancy – очікувана тривалість здорового життя;
- Freedom to make life choices – рівень свободи у виборі життєвого шляху;
- Generosity – рівень щедрості громадян;
- Perceptions of corruption (Trust) – сприйняття корупції в державних інституціях.

Цей набір даних надає достатньо можливостей для проведення глибокого аналізу, визначення основних факторів впливу на рівень щастя та побудови моделей передбачення із застосуванням методів машинного навчання.

	Country	Rank	Score	GDP	Family	Health	Freedom	Trust	Generosity
0	Switzerland	1	7.587	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678
1	Iceland	2	7.561	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630
2	Denmark	3	7.527	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139
3	Norway	4	7.522	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699
4	Canada	5	7.427	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811
...	...	...	...	...	...	...	...	...	...
151	Rwanda	152	3.334	0.35900	0.71100	0.61400	0.55500	0.41100	0.21700
152	Tanzania	153	3.231	0.47600	0.88500	0.49900	0.41700	0.14700	0.27600
153	Afghanistan	154	3.203	0.35000	0.51700	0.36100	0.00000	0.02500	0.15800
154	Central African Republic	155	3.083	0.02600	0.00000	0.10500	0.22500	0.03500	0.23500
155	South Sudan	156	2.853	0.30600	0.57500	0.29500	0.01000	0.09100	0.20200

Рис. 1. Приклад даних із датасету «World Happiness Report»

Було виконано попереднє очищення та об'єднання даних за 2015-2019 роки та проведено розвідувальний аналіз даних [2], в ході якого було створено деякі візуалізації, наведені нижче (рис. 2-4).

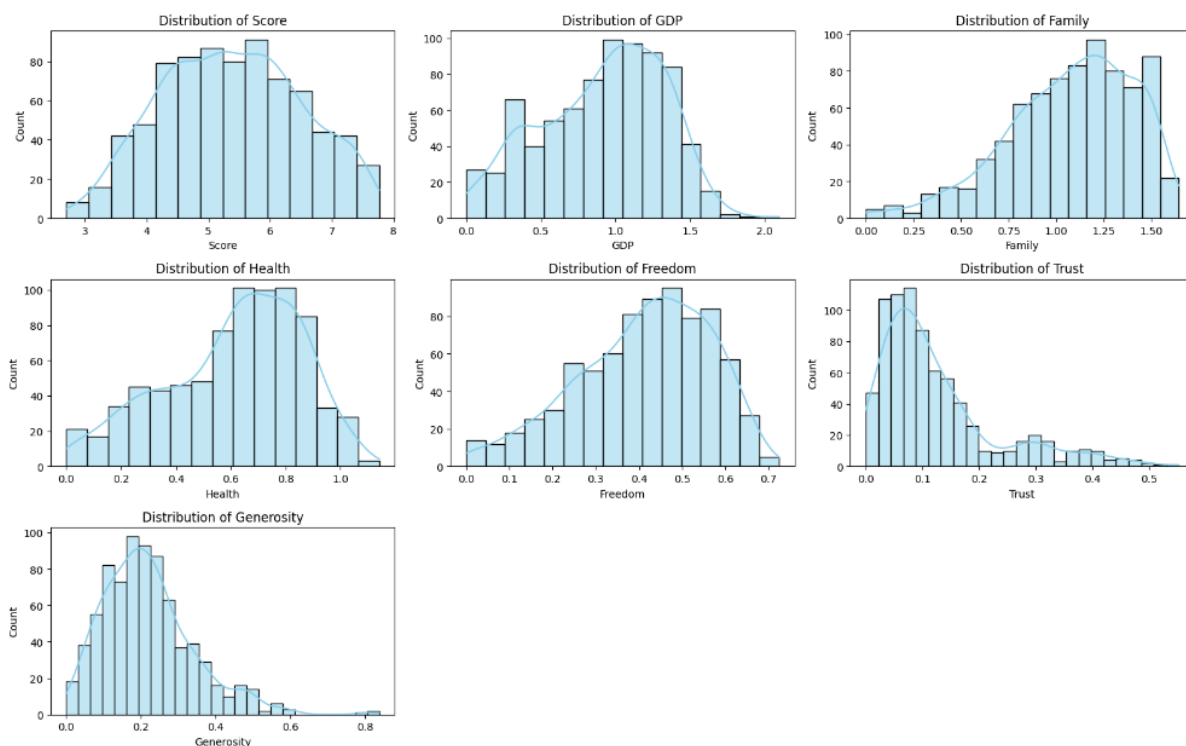


Рис. 2. Гістограма розподілу показників щастя у 2015-2019 роках

Наведені гістограми надають візуальне уявлення про розподіл ключових факторів, що впливають на щастя та економічне благополуччя країн за п'ятирічний період. Різні форми розподілів для різних показників свідчать про різницю в їхньому поширенні серед країн світу.

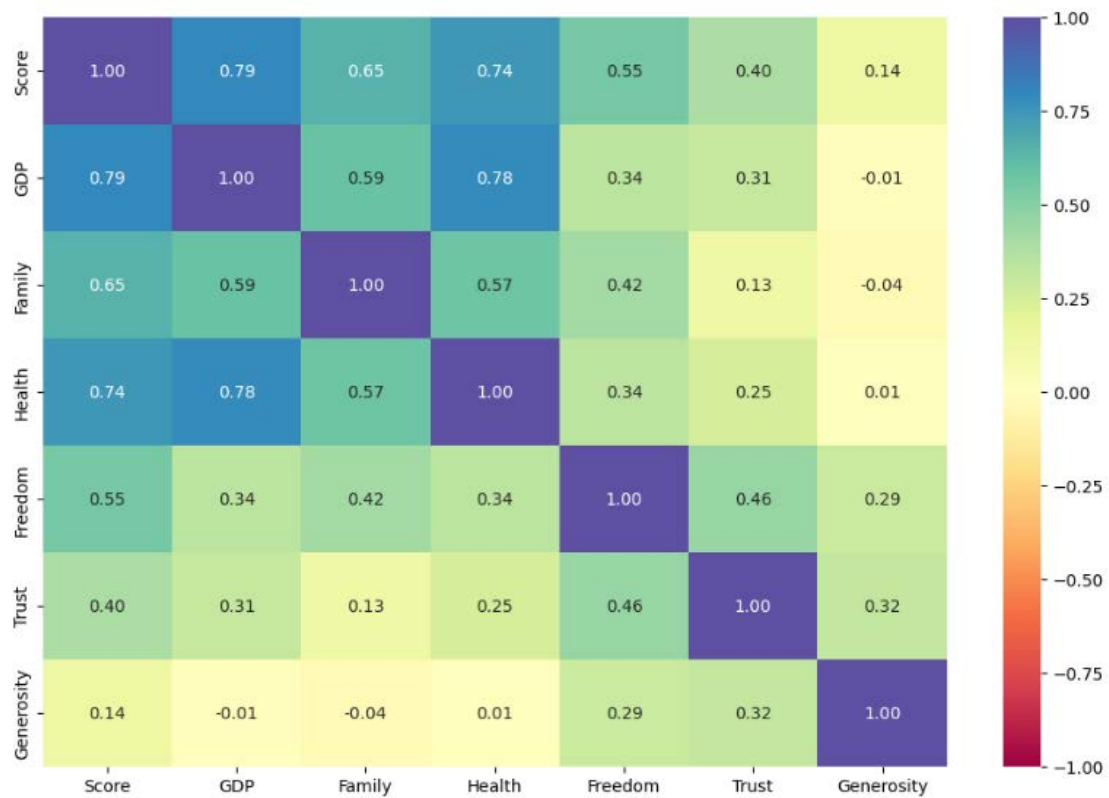


Рис. 3. Матриця кореляції

З наведеної матриці кореляції можна зробити висновок, що рівень щастя найбільше пов'язаний з економічним добробутом (ВВП), здоров'ям та соціальною підтримкою. Свобода та довіра також відіграють важливу, але меншу роль. Щедрість демонструє найслабший зв'язок з іншими досліджуваними показниками.

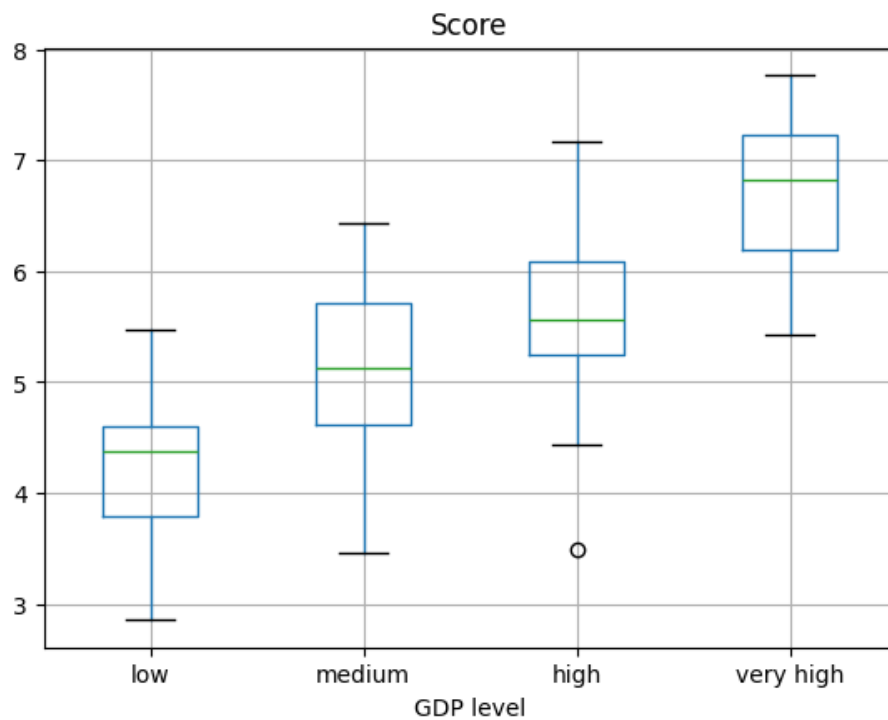


Рис. 4. Коробковий графік залежності оцінки рівня щастя від рівня економічного добробуту

На рисунку 4 видно чітку позитивну тенденцію: вищі рівні економічного добробуту (GDP level) відповідають вищим оцінкам рівня щастя (Score). Кожен коробковий графік відображає розподіл оцінок рівня щастя для країн, що належать до відповідної категорії рівня економічного добробуту («low», «medium», «high», «very high»). Такий графік наочно ілюструє, як змінюється розподіл оцінок рівня щастя в залежності від рівня економічного добробуту країн. Чітко простежується зростання як медіанного, так і міжквартильного розмаху (IQR – висота «коробки») Score зі збільшенням рівня GDP, що підтверджує позитивний зв'язок між цими показниками.

Висновки

При розвідувальному аналізі набору даних «World Happiness Report», що містить інформацію про рівень щастя та економічні показники країн за період 2015 – 2019 років, було досліджено вплив рівня економічного добробуту на оцінку рівня щастя.

Було досліджено розподіл усіх параметрів гістограм (Рис. 2). Аналіз гістограм показав, що розподіл оцінки рівня щастя (Score) достатньо рівномірний, що в майбутньому позитивно вплине на передбачення.

Також було побудовано матрицю кореляції, яка показала залежність між різними економічними та соціальними параметрами та оцінкою рівня щастя (Score). Аналіз матриці виявив помірну позитивну кореляцію між рівнем щастя та ВВП на душу населення, соціальною підтримкою та очікуваною тривалістю здорового життя.

Для більш детального вивчення впливу економічного добробуту на рівень щастя було побудовано коробковий графік (boxplot) залежності оцінки рівня щастя від рівня економічного добробуту (Рис. 4). На графіку чітко простежується тенденція – вищі рівні економічного добробуту (категорії «low», «medium», «high», «very high», сформовані на основі квантилів ВВП на душу населення) відповідають вищим значенням оцінки рівня щастя. Кожен коробковий графік відображає розподіл оцінок рівня щастя для країн з відповідним рівнем економічного добробуту, візуалізуючи медіану, квартилі, мінімальні та максимальні значення, а також потенційні викиди. Цей графік наочно демонструє позитивний зв'язок між економічним добробутом країни та середнім рівнем щастя її населення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. World Happiness Report [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/datasets/unsdsn/world-happiness/data>
2. World Happiness Report EDA & Visualization [Електронний ресурс]. – Режим доступу: <https://www.kaggle.com/code/annhaiovych/world-happiness-report-eda-visualization>
3. Pandas Getting started. 2024 [Електронний ресурс] – Режим доступу: https://pandas.pydata.org/docs/getting_started/index.html
4. Matplotlib Pyplot Documentation. 2024 [Електронний ресурс]. – Режим доступу: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html
5. Seaborn Tutorial. 2023 [Електронний ресурс]. – Режим доступу: <https://seaborn.pydata.org/tutorial.html>

Гайович Анна Михайлівна – студентка групи СА-216, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: ann.gajovich@gmail.com.

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

Haiovych Anna M. – student of Faculty of Intellectual Information Technologies and Automation, SA-21b, Vinnytsia National Technical University, Vinnytsia, e-mail: ann.gajovich@gmail.com.

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