УДК 004.94

І. Матуляк, Т. Ліхоузова, Ю. Олійник

АНАЛІЗ РІВНЯ КІБЕРРИЗИКІВ В КРАЇНАХ СВІТУ

Анотація: Робота присвячена кластеризації та класифікації країн за рівнем кіберризиків на основі відібраних індексів кібербезпеки, економічних та технологічних факторів. Результати роботи можуть бути корисними для організацій, що займаються моніторингом кіберзагроз, формуванням політик у сфері кібербезпеки, технологічних компаній, а також для аналітичних центрів, які оцінюють рівень кіберризиків різних країн з урахуванням економічних та технологічних факторів.

Ключові слова: інтелектуальний аналіз даних, модель кластеризації, модель класифікації.

Вступ

У сучасних умовах глобальної цифровізації та інтеграції інформаційних технологій у всі сфери суспільного життя, питання кібербезпеки набуває особливої актуальності. Зростання обсягів передавання, обробки та зберігання даних, активне впровадження хмарних сервісів, інтернету речей та автоматизованих систем управління призводить до зростання уразливості державних установ, комерційних структур та окремих громадян до кіберзагроз. Як наслідок, виникає потреба у створенні ефективних інструментів виявлення, аналізу та протидії таким загрозам.

Матеріали та методи

Одним із ключових напрямів у контексті протидії загрозам, зокрема шляхом розробки спеціалізованих систем збирання, обробки та зберігання інформації про кіберінциденти, є побудова централізованого сховища даних, що забезпечує не лише акумуляцію релевантної інформації, а й надає аналітичні можливості для виявлення закономірностей, прогнозування атак і формування запобіжних заходів. Необхідність проєктування та впровадження такого сховища зумовлена стрімким зростанням кількості кіберінцидентів та ускладненням їхніх форм, що становить реальну загрозу як для державних інституцій, так і для приватного сектору.

Для аналізу та моделювання було обрано 3 джерела відкритих даних на сайті https://www.kaggle.com// та https://data.worldbank.org/, що включають в себе:

− Рейтинг кібербезпеки за регіонами: https://www.kaggle.com/datasets/katerynameleshenko/cyber-security-indexes

− Набір даних, що містить статистику про основні кіберзагрози у різних країнах за період 2015-2024 років. Включає інформацію про типи атак, кількість інцидентів,

© І. Матуляк, Т. Ліхоузова, Ю. Олійник

 ISSN 1560-8956

географічний розподіл загроз та їхні наслідки: https://www.kaggle.com/datasets/atharvasoundankar/global-cybersecurity-threats-2015-2024

− Набір даних, що містить світові індикатори розвитку: https://databank.worldbank.org/source/world-development-indicators#

У моделі сховища за типом сніжинка спроектовано три таблиці фактів та сім таблиць вимірів.

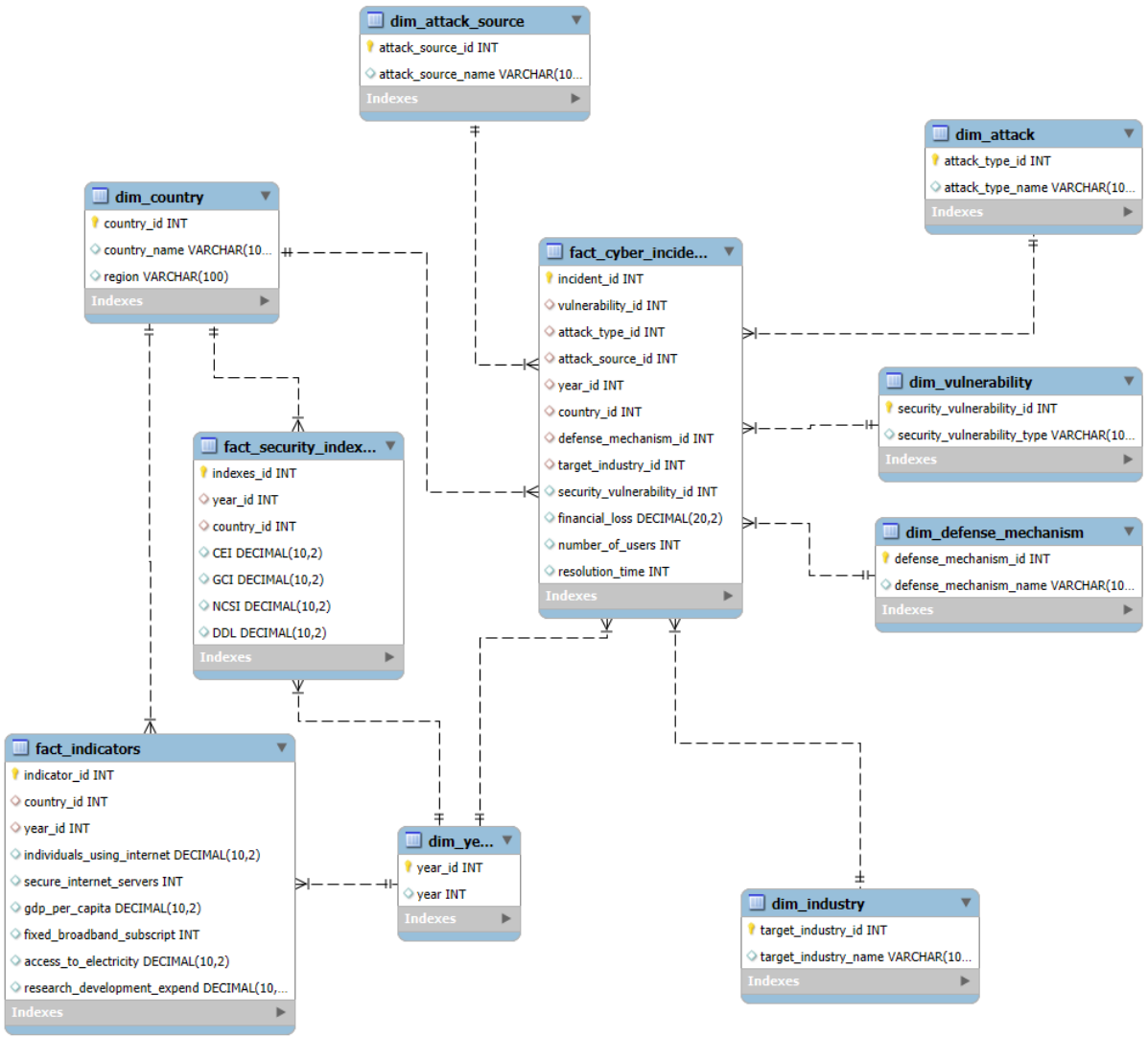

Рисунок 1. Модель сховища за типом сніжинка

Далі було виконано повний цикл об’єднання, очищення та уніфікації даних із різних джерел, що дозволило створити цілісний, структуровано узгоджений та аналітично придатний датасет. Завдяки застосованим методам обробки вдалося мінімізувати кількість пропущених значень, виявити та скоригувати потенційні аномалії, а також забезпечити узгодженість форматів і назв змінних.

Для дослідження кібербезпеки та виявлення закономірностей у даних було обрано методи машинного навчання, які дозволяють аналізувати складні залежності між різними факторами, такими як індекси кібербезпеки, соціально-економічні показники

та рівень кіберзагроз. Вибір методів ґрунтується на специфіці даних, які включають як кількісні, так і якісні характеристики, а також на необхідності виявляти кластери країн з подібними ризиками та прогнозувати потенційні загрози.

Для кластеризації даних було обрано алгоритми K-Means [1], DBSCAN [2] та Agglomerative Clustering [3]. K-Means є ефективним для розділення країн на групи за схожими показниками кібербезпеки, оскільки він дозволяє визначити чіткі центроїди кластерів на основі евклідової відстані. Цей метод особливо корисний для виявлення глобальних тенденцій у різних регіонах. DBSCAN обрано для виявлення аномалій та країн із нестандартними профілями кібербезпеки, оскільки він не вимагає попереднього задання кількості кластерів і здатний знаходити складні форми розподілу даних. Agglomerative Clustering дозволяє будувати ієрархічну структуру кластерів, що дає змогу аналізувати вкладені групи країн із різним рівнем деталізації.

Для задач класифікації, таких як прогнозування рівня кіберзагроз на основі економічних і технологічних показників, було обрано алгоритми K-Nearest Neighbors (KNN) [4], Random Forest [5] та Gradient Boosting [6]. KNN є простим, але ефективним методом для класифікації на основі схожості з найближчими сусідами, що дозволяє враховувати локальні закономірності у даних. Random Forest обрано через його здатність обробляти велику кількість ознак із різною значимістю, а також через стійкість до перенавчання. Цей метод особливо корисний для аналізу складних залежностей між індексами кібербезпеки та зовнішніми факторами. Gradient Boosting дозволяє будувати точні моделі за рахунок послідовного покращення прогнозів, що важливо для передбачення рідкісних, але критичних кіберінцидентів.

Вибір цих методів обумовлений їхньою здатністю обробляти великий обсяг різнорідних даних, включаючи часові ряди, категоріальні ознаки та числові показники. Кластеризаційні алгоритми допоможуть виявити групи країн із подібними профілями кібербезпеки, що може бути використано для розробки регіональних стратегій захисту. Методи класифікації дозволять прогнозувати рівень загроз на основі історичних даних, що є ключовим для запобігання кібератакам. Усі обрані алгоритми реалізовані у бібліотеці scikit-learn, що забезпечує їхню ефективність та масштабованість [7].

Результати та обговорення

Моделі для кластерного аналізу

Для моделювання обрано найбільш інформативні індекси кібербезпеки (GCI, CEI, NCSI, DDL), економічні та інфраструктурні показники (GDP per Capita, Internet Users, Secure Servers, Broadband Subscriptions, Electricity Access, R&D Expenditures). Ознаки було обрано на основі кореляційного аналізу з цілями: кількість атак, фінансові втрати, кількість постраждалих користувачів. Ознаки з низькою варіативністю або відсутнім внеском з моделі виключено(рис. 2 – 3).

 ISSN 1560-8956

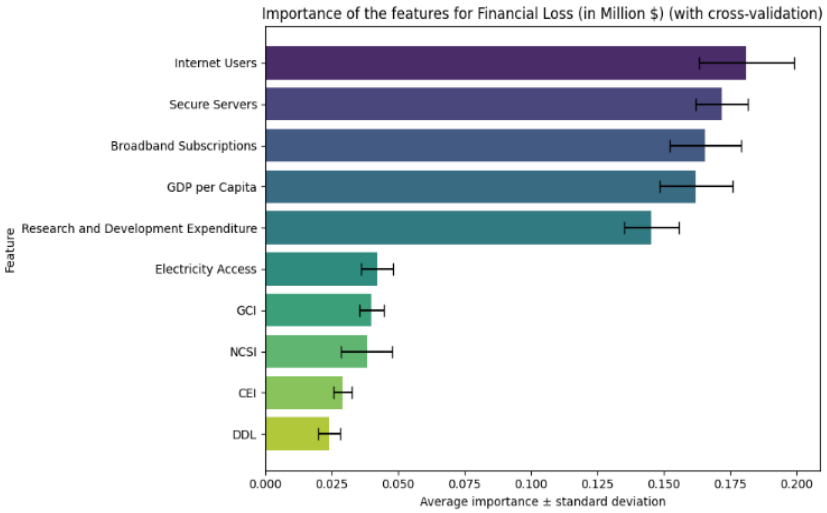

Рисунок 2. Значущість ознак по відношенню до кількості втрат від атак.

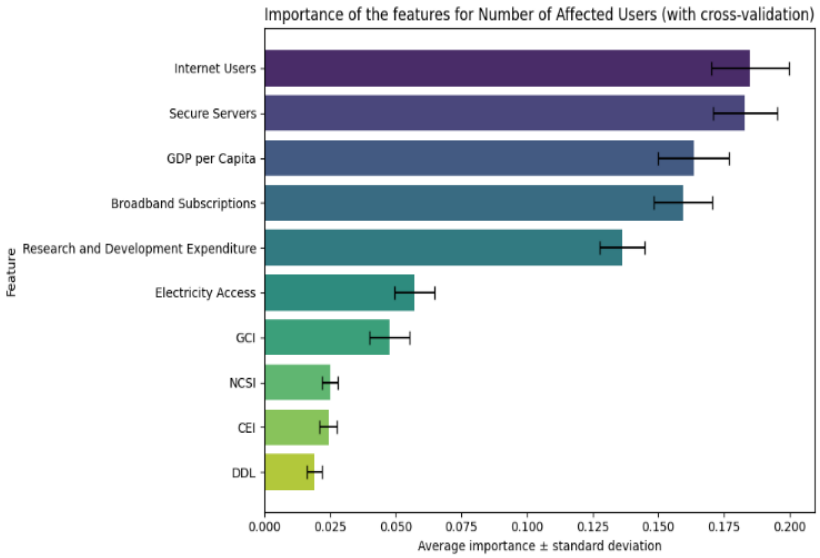

Рисунок 3. Значущість ознак по відношенню
до кількості постраждалих від атак користувачів

Проведений аналіз дозволив визначити ключові фактори, що найбільше впливають на різні аспекти кіберзагроз. У випадку фінансових втрат (Financial Loss) найвищу залежність виявлено для таких показників, як кількість користувачів Інтернету, кількість захищених серверів, рівень широкосмугового підключення та ВВП на душу населення. Серед індексів кібербезпеки найбільший вплив мають GCI (Global Cybersecurity Index) та NCSI (National Cybersecurity Commitment Index).

Аналогічна тенденція спостерігається і для кількості постраждалих користувачів (Number of Affected Users), де вирішальними залишаються Internet Users, Secure Servers і GDP per Capita, а також згадані індекси GCI та NCSI.

Щодо загальної кількості кібератак, то тут найсильніший вплив мають саме GCI та DDL (Digital Development Level), що підтверджує важливість рівня цифрової інфраструктури та зрілості системи кіберзахисту як чинників ризику або виявлення атак.

Отже, обрані фактори для використання в подальшому аналізі: Internet Users, Secure Servers, Broadband Subscriptions, GDP per Capita, GCI, NCSI, DDL.

Перед тренуванням моделей виконано стандартизацію числових показників методом Z-нормалізації (StandardScaler) [16], що є критично важливим для моделей, чутливих до масштабу (K-Means, KNN, DBSCAN).

На рисунку 4 наведено результати кластеризації країн по рівнях кіберризиків за GCI та DDL. Для інших комбінацій обраних предикторів результати доволі схожі.

Для оцінки ефективності кластеризації країн за рівнем кіберризику було використано три різні підходи: K-Means, DBSCAN та Agglomerative Clustering. Результати кластеризації узагальнено в таблиці 1, де наведено кількість утворених кластерів, частку "шуму" для DBSCAN, а також найкраще досягнуте значення силуетного коефіцієнта, що є одним з основних показників якості кластеризації.

Таблиця 1.

Method	**Number of Clusters**	**% Noise (for DBSCAN)**	**Best Silhouette Score**
K-Means	6	-	0.88
DBSCAN	5	1.2%	0.776
Agglomerative Clustering	6	-	0.85

Метод K-Means поділив країни на 6 кластерів, продемонструвавши найвищу якість кластеризації за силуетним коефіцієнтом (0.88). Це свідчить про те, що країни в межах кожного кластера мають високий ступінь подібності, а відстань до сусідніх кластерів – суттєва. Така класифікація дозволяє виявити як країни з високим рівнем кіберзахисту, так і держави, що характеризуються критичним рівнем ризику, або займають проміжну позицію. Створення шести кластерів забезпечує деталізоване багаторівневе групування, яке зручно використовувати для подальшого моделювання чи формування політик безпеки.

Метод DBSCAN, який базується на щільності, сформував 5 кластерів та класифікував 1.2% країн як "шум" – тобто країни, які не належать до жодного з основних кластерів через свою аномальну поведінку або унікальні поєднання характеристик. Значення силуетного коефіцієнта – 0.776 – свідчить про прийнятну, але нижчу якість кластеризації порівняно з іншими методами. Основна перевага DBSCAN полягає в тому, що він дозволяє виявити нетипові країни або приховані структури, не потребуючи попереднього задання кількості кластерів. Проте такий підхід є менш стабільним і складнішим для інтерпретації.

 ISSN 1560-8956

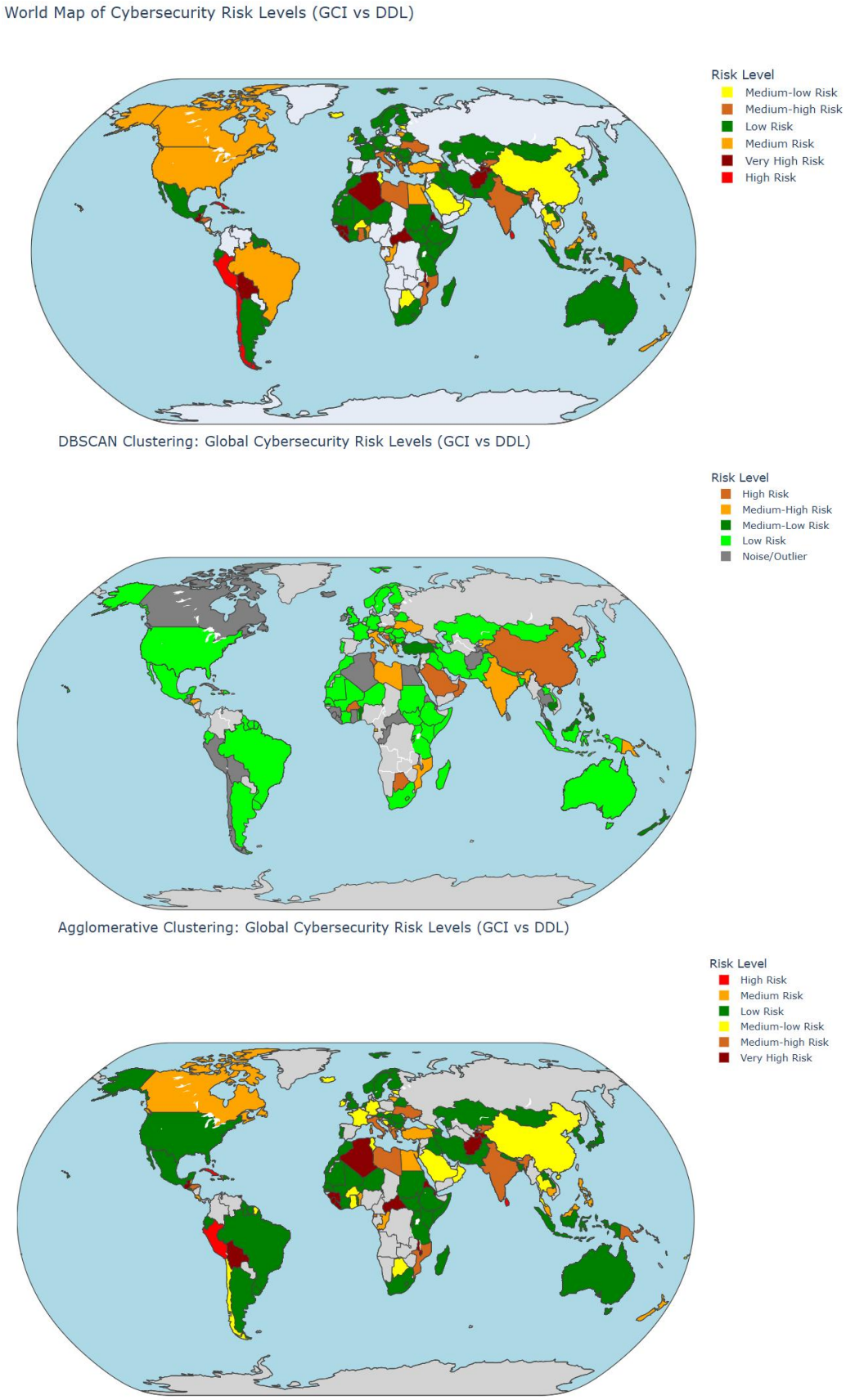

Рисунок 4. Результати кластеризації від 3-ох моделей

Метод Agglomerative Clustering (ієрархічна кластеризація) також сформував 6 кластерів, аналогічно до K-Means, і показав високий силуетний коефіцієнт – 0.85. Цей результат є майже ідентичним до результатів K-Means за структурою, що свідчить про стійкість і узгодженість кластерної структури даних незалежно від обраного алгоритму. Agglomerative Clustering дозволяє візуально відстежити процес об'єднання країн у кластери (через дендрограми), що зручно для додаткового аналізу близькості між об'єктами.

Загалом, метод K-Means визнано найоптимальнішим для кластеризації країн за рівнями кіберризику, враховуючи співвідношення точності, стабільності та простоти результатів.

Моделі для класифікації

Було проведено порівняльний аналіз трьох моделей машинного навчання: KNN, Random Forest та Gradient Boosting, які були застосовані для класифікації країн за рівнем кіберризику. Критеріями порівняння обрано точність (accuracy), precision, recall та F1-середнє, що дозволяє оцінити якість класифікації для кожної з категорій (низький, середній та високий рівні ризику).

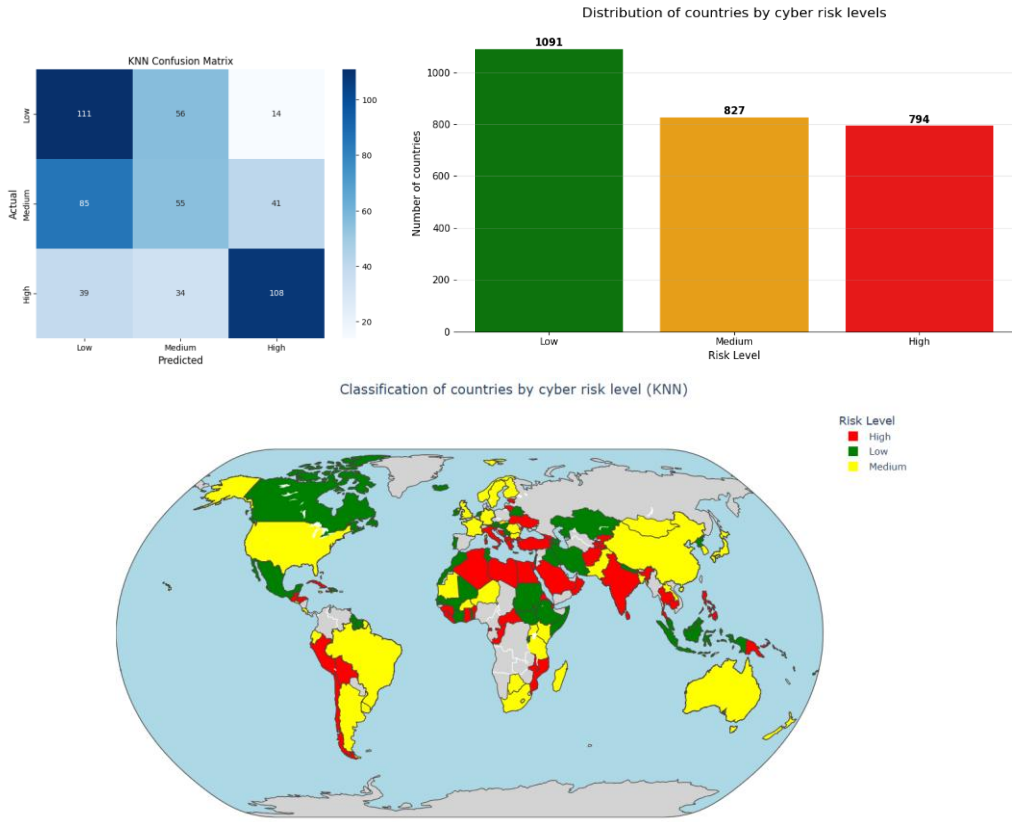

Рисунок 5. Результати класифікації країн за рівнями кіберризиків моделлю KNN

У результаті класифікації рівнів кіберризику для 145 країн за трьома різними алгоритмами виявлено суттєві розбіжності між оцінками, що свідчить про чутливість моделі до обраного методу. Із загальної кількості у 2712 записів лише 27 країн отримали однаковий рівень ризику за всіма трьома методами, що становить менш як

 ISSN 1560-8956

п’яту частину загальної вибірки. У 116 країнах зафіксовано часткову розбіжність (два з трьох методів не збігаються у класифікації), а в 2 країнах – повну невідповідність між усіма трьома оцінками.

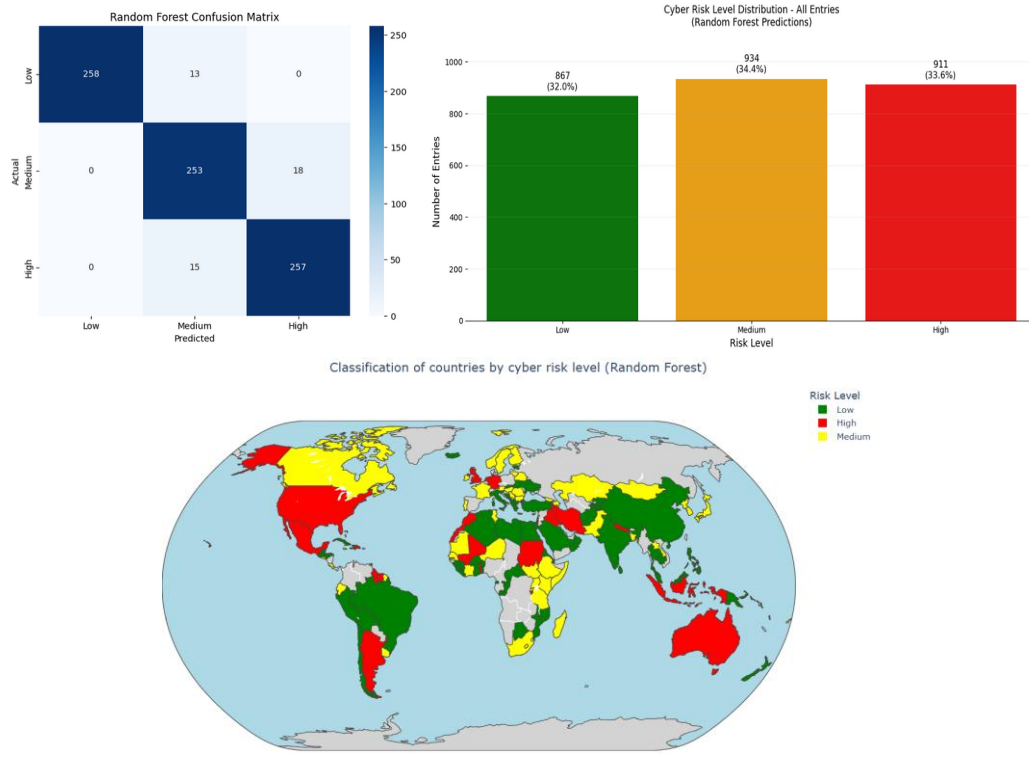

Рисунок 6. Результати класифікації країн моделлю Random Forest

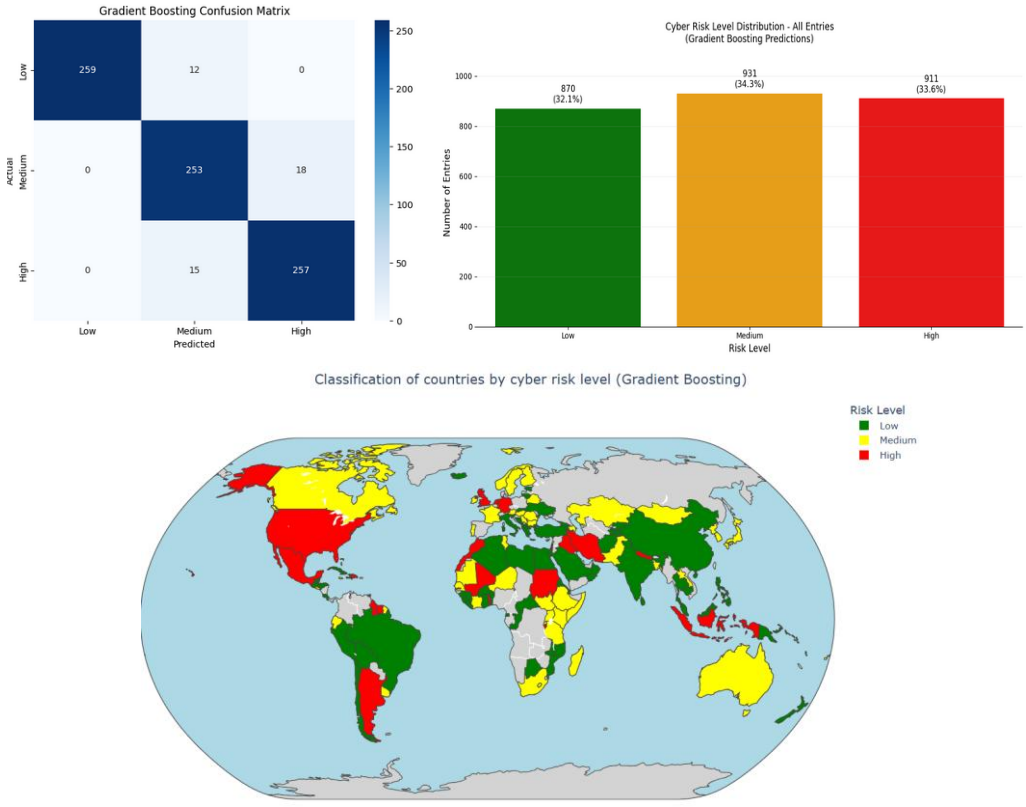

Рисунок 7. Результати класифікації країн моделлю Gradient Boosting

Найяскравішими прикладами повної невідповідності стали Грузія (Georgia) та Литва (Lithuania), які були класифіковані як країни з високим ризиком за KNN, середнім – за Random Forest та низьким – за Gradient Boosting. Це свідчить про те, що вхідні дані для цих країн є неоднозначними або перебувають на межі між класами, що призводить до різної інтерпретації залежно від обраного алгоритму.

У багатьох випадках KNN демонстрував схильність до класифікації країн як таких, що мають високий рівень кіберризику, навіть тоді, коли інші методи (особливо Gradient Boosting) оцінювали їхній ризик як низький. Така поведінка пояснюється тим, що KNN базується на близькості до сусідніх точок у багатовимірному просторі ознак і не враховує складних внутрішніх взаємозв'язків між змінними. Наприклад, такі країни, як Brazil, India, Samoa, Cambodia та багато інших, були класифіковані як "High Risk" за KNN, у той час як Gradient Boosting і Random Forest відносили їх до групи з низьким або середнім рівнем ризику.

Водночас між Random Forest і Gradient Boosting спостерігалася вища узгодженість. Ці два методи, обидва базовані на ансамблях дерев рішень, часто давали однакові оцінки ризику, особливо для таких країн, як USA, UK, Germany, Netherlands, Israel, Indonesia та ін. Це дозволяє зробити висновок, що ансамблеві методи, які враховують складні патерни у даних, є більш надійними для подібного типу класифікації.

Gradient Boosting, який продемонстрував найвищу точність, показав себе як найстабільніший та найнадійніший метод класифікації у цьому дослідженні. Він виявляє закономірності глибшого рівня, ніж KNN, і водночас менш схильний до переобучення порівняно з Random Forest. Саме оцінки Gradient Boosting були прийняті як фінальні для визначення рівнів кіберризику по країнах.

Таким чином, аналіз результатів класифікації свідчить про важливість вибору правильного алгоритму машинного навчання у завданнях, пов'язаних з оцінкою загроз. Метод Gradient Boosting варто використовувати як базовий інструмент для формування політик кібербезпеки, прогнозування загроз та міжнародного порівняння ризиків, адже він забезпечує як високу точність, так і стійкість до неоднозначностей у даних.

Висновки

На основі детального опису та аналізу предметної області було виконано кластеризацію та класифікацію країн світу за рівнем кіберризиків із використанням актуальних індексів кібербезпеки, економічних і технологічних показників. За допомогою моделей класифікації K-Nearest Neighbors, Random Forest та Gradient Boosting здійснено прогнозування рівня ризику для кожної країни. Найкращий результат показала модель Gradient Boosting, яка досягла точності близько 94.47%. Результати класифікації підтверджені обчисленням точності та інших метрик оцінювання, а також візуальним та порівняльним аналізом розбіжностей між методами.

 ISSN 1560-8956

Кластеризація за допомогою алгоритмів K-Means, DBSCAN та Agglomerative Clustering дозволила виявити три основні групи країн: з високим, середнім та низьким рівнем кіберризиків.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Zakeri A., Qin C., Gebremariam B.* Cybersecurity Clustering Using K-Means [Електронний ресурс] // arXiv preprint. – 2024. – arXiv:2402.01061. – Режим доступу: https://arxiv.org/pdf/2402.01061.

2. scikit-learn developers. sklearn.cluster.DBSCAN [Електронний ресурс]. – Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.

3. scikit-learn developers. sklearn.cluster.AgglomerativeClustering [Електронний ресурс]. – Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html.

4. *Altman N.* An Introduction to K-Nearest Neighbors [Електронний ресурс] // arXiv preprint. – 2017. – arXiv:1706.03113. – Режим доступу: https://arxiv.org/pdf/1706.03113.

5. *Breiman L.* Random Forests [Електронний ресурс] // arXiv preprint. – 2011. – arXiv:1109.2378. – Режим доступу: https://arxiv.org/pdf/1109.2378.

6. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System [Електронний ресурс] // arXiv preprint. – 2020. – arXiv:2012.06047. – Режим доступу: https://arxiv.org/pdf/2012.06047.

7. scikit-learn developers [Електронний ресурс]. – Режим доступу: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.