9. Україна : перспективи розвитку (Консенсус-прогноз): Міністерство економічного розвитку і торгівлі України ; ПРООН. — 2017. — Вип. 43. — 33 с.
10. Україна: перспективи розвитку (Консенсус-прогноз) : Міністерство економічного розвитку і торгівлі України ; ПРООН. — 2012. — Вип. 30. — 34 с.

1.23. Використання інструментів SAS ENTERPRISE MINER для ідентифікації терористичного угрупування, відповідального за терористичний акт

Постановка проблеми. На рубежі XX–XXI століть тероризм зараховують до числа найбільш небезпечних і важко прогнозованих явищ, він набуває все більш різноманітних форм та загрозливих масштабів. На даний час, за експертними оцінками, у світі діє понад 500 терористичних організацій і груп різної спрямованості. За своєю географією список охоплює, по суті, всю планету [1]. Зважаючи на зростання світового рівня терористичної загрози, а також на воєнні дії, які ведуться на території України питання підвищення ефективності заходів боротьби з тероризмом стає надзвичайно актуальним [2].

Терористичні атаки є найбільшим викликом для людства та зумовлюють націленість світової спільноти на боротьбу з ними. Найефективнішим інструментом контртероризму є притягнення винних до відповідальності за вчинений злочин. Однак результативність даного інструменту прямо залежить від швидкої ідентифікації організаторів та виконавців терористичних актів. Так, з 79534 терористичних актів здійснених з 2005 по 2015 роки у 44672 випадках не було визначено особу чи угрупування, яке вчинило цей акт [3]. З огляду на це, актуальність досліджень спрямованих на класифікацію терористичних актів за терористичними угрупуваннями, відповідальними за їх організацію та вчинення не викликає сумніву. З огляду на це, метою даного дослідження є розгляд можливості використання інструментів *SAS Enterprise Miner* для ідентифікації терористичних угрупувань, відповідальних за вчинення терористичних актів, результати якого можуть бути використані для звуження кола підозрюваних, а отже підвищення ефективності боротьби з тероризмом в цілому.

Аналіз останніх досліджень і публікацій. Теоретичні аспекти та практичний досвід щодо використання інструментів інтелектуального аналізу даних в контртерористичній діяльності розглядаються як у світовій так і у вітчизняній науковій літературі. Однак питання ідентифікації терористичного угрупування, відповідального за терористичний акт є досить специфічною предметною областю, яка мало представлена в працях вітчизняних науковців.

Тож *метою* статті є розгляд можливості використання інструментів SAS Enterprise Miner для побудови класифікаційної моделі для ідентифікації особи чи терористичного угрупування, які вчинили терористичний акт задля притягнення їх до відповідальності.

Виклад основного матеріалу. Для проведення дослідження була використана Global Terrorism Database (*GDT*), яка належить National Consortium for the Study of Terrorism and Responses to Terrorism (*START*), і містить світову інформацію по терористичним актам з 1970 року.

Із бази даних було сформовано вибірку за період 2005-2015 рр. із 26353 спостережень, в яких було точно ідентифіковано терористичну групу, що несе відповідальність за скоєний злочин. В подальшому із неї було виключено записи по спостереженням, які були скоєні терористичними групами із порівняно малою кількістю інцидентів.

Незалежними змінними для побудови моделі обрано ті, які є відомими після вчинення терористичного акту. Перелік, зміст, роль та тип змінних представлено в табл. 1.

Таблиця 1

Опис вхідних даних

Змінна	Зміст	Тип	Роль
gname	**Назва терористичного угрупування, яке несе відповідальність за терористичний акт**	**nominal**	**цільова**
attactype1	Метод нападу, який був використаний під час терористичного акту	nominal	вхідна
crit1	Належність цілі акту до політичних, економічних, релігійних чи соціальних цілей	binary	вхідна
crit2	Відображає намір примушування, залякування чи опублікування для більших аудиторій	binary	вхідна
crit3	Відображає вихід акту за межі міжнародного гуманітарного права	binary	вхідна
multiple	Вказує на зв'язок терористичного акту з іншими терористичними актами	binary	вхідна
nkill	Кількість загиблих, включаючи жертв та терористів	interval	вхідна
nwound	Кількість не смертельно поранених жертв та терористів	interval	вхідна
property	Вказує на наявність матеріального збитку від нападу	nominal	вхідна
region	Вказує на регіон, в якому був вчинений напад	nominal	вхідна
success	Вказує на успішність терористичної атаки, виходячи з матеріальних наслідків	binary	вхідна
suicide	Вказує на те, що злочинець не збирався вижити після атаки	binary	вхідна
targtype1	Загальний тип цілі чи мішені (бізнес, уряд, армія тощо)	nominal	вхідна
targsubtype1	Точний тип цілі чи мішені (банк, завод, готель, президент, парламент тощо)	nominal	вхідна

Змінна	Зміст	Тип	Роль
weaptype	Тип зброї, використаної терористами	nominal	вхідна

Вхідний набір даних задовольняє наступним вимогам:

– Кожен запис набору даних асоціюється з одним із класів — змінна *gname* є міткою класу.

– Класи є дискретними — кожне спостереження вибірки даних однозначно відноситься до конкретного терористичного угрупування.

Кількість класів значно менше кількості записів досліджуваного набору даних — результуючий набір даних містить 26353 спостережень, які розподілені за 34 терористичним угрупуванням.

У випадку побудови класифікаційних моделей, в яких цільова змінна приймає дискретні значення, найбільш популярним і потужним інструментом інтелектуального аналізу даних (*Data Mining*) є дерева рішень (*Decision Trees*).

В інтелектуальному аналізі даних дерева рішень можуть бути використані в якості математичних і обчислювальних методів, щоб допомогти описати, класифікувати й узагальнити набори даних, які можуть бути записані таким чином:

$$(x, Y) = (x_1, x_2, x_3, \ldots, x_k, Y). \quad (1)$$

Залежна змінна Y для даного дослідження є цільовою змінною *gname*, яку необхідно класифікувати. Вектор x складається з вхідних змінних $x_1, x_2, x_3, \ldots, x_k$ – для даного дослідження $k = 14$.

В основі роботи дерев рішень лежить процес рекурсивної розбивки вхідної множини спостережень або об’єктів на підмножини, асоційовані з класами.

Переваги методу побудови дерева рішень:

– класифікаційна модель, представлена у вигляді дерева рішень, є інтуїтивною і спрощує розуміння розв'язуваної задачі;

– не потрібно апріорних припущень про вид залежності між досліджуваними даними;

– дерева рішень стійкі до «прокляття розмірності»;

– дерева рішень стійкі до викидів в просторі ознак;

– працюють як з числовими так і з категоріальними даними, тобто не потрібно перекодувати категоріальні змінні;

– не потрібна підстановка пропущених значень;

– швидке навчання [4].

Алгоритм конструювання дерева рішень:

Eman 1. «Побудова» дерева (*tree building*) – вирішуються питання вибору критерію розщеплення й зупинки навчання (якщо це передбачене алгоритмом).

Eman 2. «Скорочення» дерева (*tree pruning*) – вирішується питання відсікання деяких його гілок [5].

На рис. 1 наведено діаграму проекту в *SAS*. Першим блоком діаграми виступає уся сукупність даних – класифіковані терористичні акти. У другому блоці діаграми процесу використовується інструмент Фільтрація (*Filtering*), за допомогою якого було виключено з дослідження рідкісні значення вхідних змінних, які зустрічаються менш ніж у 1% випадків. Всього було відфільтровано 2309 значень. Третій етап роботи – Розподіл даних (*Data Partition*), за допомогою якого уся вхідна сукупність даних рандомно ділиться на 2 частини: 50% – тренувальні дані (*training*), на яких модель буде будуватись, 50% – дані, на яких модель буде перевірятись (*validation*). Четвертий етап роботи – побудова і оптимізація моделі дерева рішень: автономне дерево рішень (*Autonomous Decision Tree*) та інтерактивне дерево рішень (*Interactive Decision Tree*). П'ятий етап – використання блоку Порівняння моделей (*Model Comparison*) для порівняння створених моделей на валідаційному наборі даних.

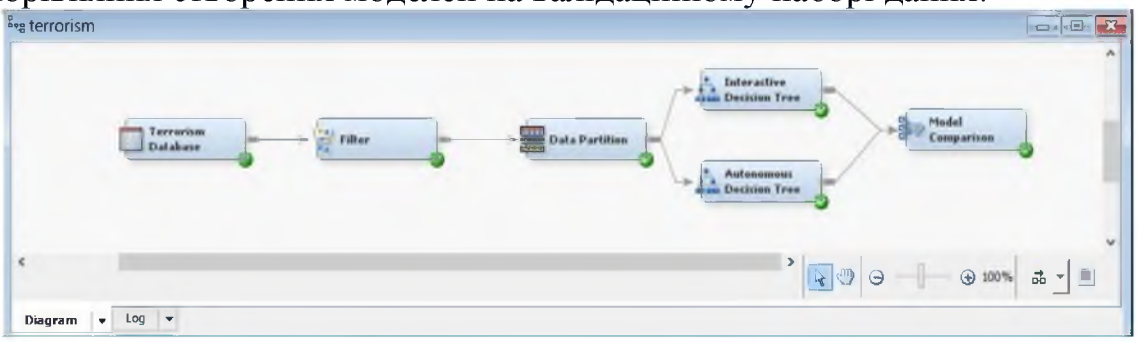


Рисунок 1 – Діаграма процесу побудови моделей в пакеті *SAS Enterprise Miner*

При побудові інтерактивного дерева рішень було зроблено наступні початкові налаштування інструменту *Decision Tree*:

– вид розбиття – множинне (*multi-way*) – змінюємо максимальну кількість можливих гілок на 7, щоб можна було проводити класифікацію терористичних атак з прив'язкою до конкретного регіону;

– змінюємо значення *Missing values* зі значення за замовчуванням на *largest branch*, щоб пропущені значення відносились до більшої гілки;

– критерій розбиття – χ^2;

– відключаємо поправку Бонферроні (*Bonferroni Adjustment*), це робить метод χ^2 більш дієвим для змінних з великою кількістю рівнів ряду.

Перша частина алгоритму побудови інтерактивного дерева рішень — пошук розгалужень, яка починається з вибору вхідної змінної для розбиття наявних навчальних даних. У ході процесу алгоритм повинен знайти такий критерій розщеплення, щоб розбити множину на підмножини, які б асоціювалися з даним вузлом перевірки [6].

Для великих наборів даних якість розділення описується значенням

$$\log worth = -\log_{10}(\chi^2 \ for \ p-value). \qquad (2)$$

Як мінімум для однієї змінної $\log worth$ повинне перевищувати порогове значення, для того щоб по даній вхідній змінній відбулося розгалуження. За замовчуванням порогове значення відповідає ймовірності $\chi^2 = 0{,}2$ або приблизному значенню $\log worth = 0{,}7$.

Враховуючи вищевикладене, кореневий вузол було розбито за змінною *region*, значення *logworth* для якої рівне 12478,042 (рис. 2).

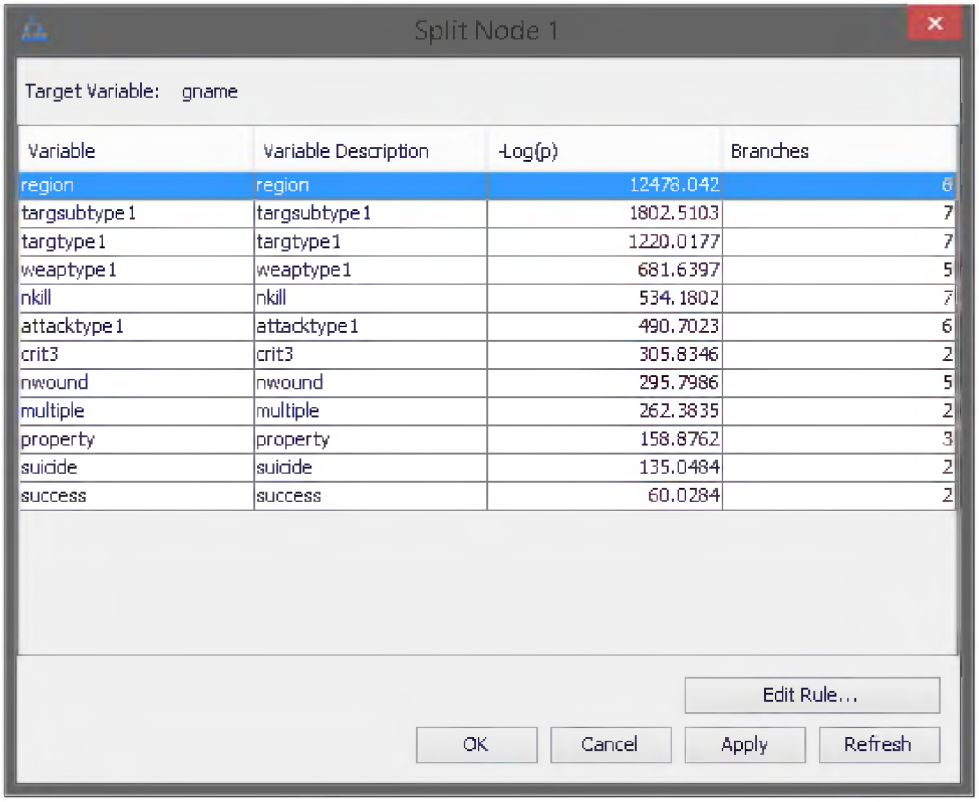


Рисунок 2 – Процес пошуку розгалужень

Результатом першого розгалуження було отримано шість гілок, які відповідають одному із шести регіонів, що є у сформованій вибірці вхідних даних. Провівши розгалуження по кожному з утворених вузлів згідно правила максимальності значення *logworth*, було отримано дерево, зображене на рис. 3-4.

Якість класифікаційної моделі, побудованої за допомогою дерева рішень, характеризується двома основними ознаками:

– *Точність класифікації* – відношення об'єктів, правильно класифікованих в процесі навчання, до загальної кількості об'єктів набору даних, які використовувались для навчання.

– *Помилка* – відношення об’єктів, неправильно класифікованих в процесі навчання, до загальної кількості об’єктів набору даних, які використовувались для навчання [5].

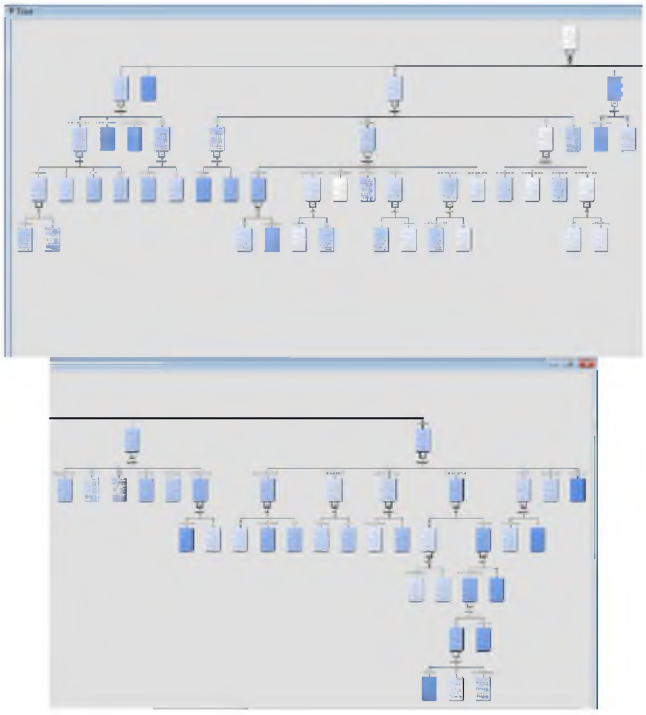


Рисунок 3 – Максимальне вирощене інтерактивне дерево рішень

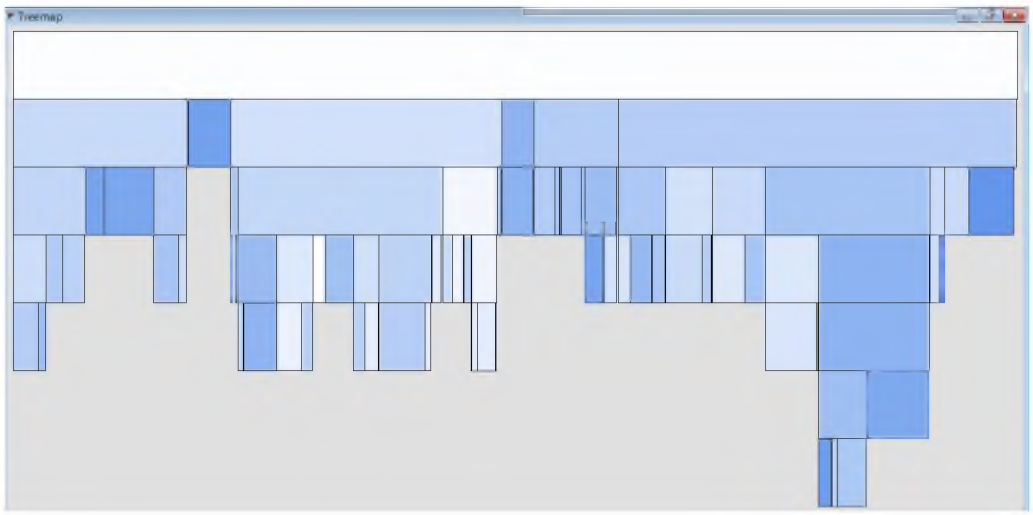


Рисунок 4 – Максимальне вирощене інтерактивне дерево рішень – *Treemap*

Коефіцієнт помилкової класифікації (*Misclassification Rate*) для тренувального і валідаційного наборів даних має спадну тенденцію. Це говорить про те, що зі збільшенням кількості гілок дерева коефіцієнт помилкової класифікації зменшується, тобто, дерево краще класифікує дані. На 34-му кроці цей показник досягає свого мінімального значення для валідаційного набору, отже подальше нарощування кількості голок не

є доцільним (рис. 5). Таким чином, у якості оптимального варіанту за замовчуванням було обрано дерево з 34-ма гілками розгалужень.

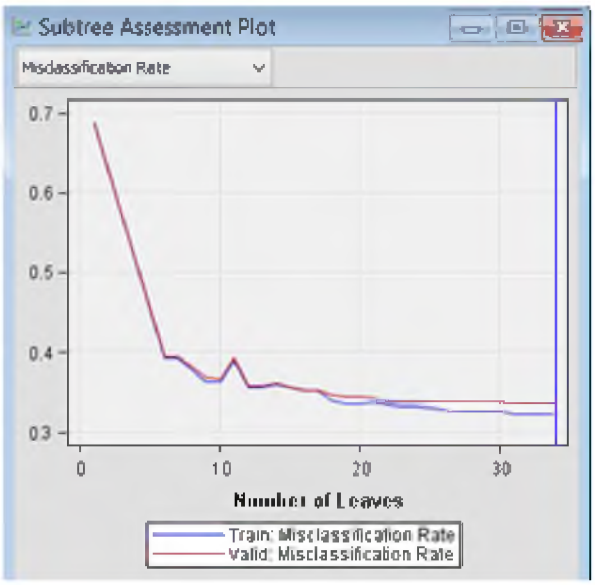


Рисунок 5 – Графік зміни значень *Misclassification Rate* після скорочення дерева рішень

Розглянемо прогнозовані значення моделі (рис. 6).

Leaf Statistics Table

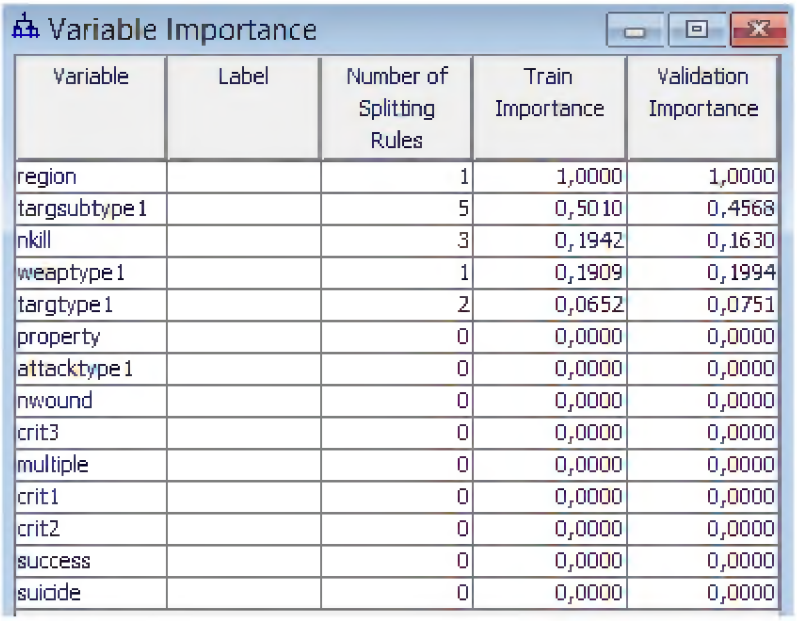
No...	Predicted Target Value	Train Frequency	Train % Corr...	Validation Freque...	Validation % Corr...
63	BALOCH REPUBLICAN ARMY (BRA)	66	89,39	51	94,12
31	TALIBAN	544	91,54	564	92,55
4	REVOLUTIONARY ARMED FORCES OF CO	511	84,74	511	84,74
11	AL-SHABAAB	600	80,67	611	79,71
10	BOKO HARAM	222	82,88	233	76,39
17	DONETSK PEOPLE'S REPUBLIC	381	72,44	380	72,63
61	TALIBAN	1329	72,31	1243	71,52
22	NEW PEOPLE'S ARMY (NPA)	237	60,34	242	69,42
18	MUSLIM FUNDAMENTALISTS	5	40,00	6	66,67
13	PALESTINIAN EXTREMISTS	94	59,57	104	65,38
24	NEW PEOPLE'S ARMY (NPA)	381	65,62	394	62,44
34	AL-SHABAAB	209	55,98	232	61,21
25	TALIBAN	574	57,67	503	58,45
16	KURDISTAN WORKERS' PARTY (PKK)	49	55,10	60	58,33
65	BOKO HARAM	87	64,37	87	57,47
35	BOKO HARAM	258	50,78	243	55,97

Рисунок 6 – Вигляд вікна *Leaf Statistic table* – Прогнозовані значення класифікаційного дерева рішень

Побудована модель може з точністю 94,12% класифікувати терористичний акт, як злочин, вчинений Республіканською армією Белуджистану, з точністю 92,55% — як злочин, вчинений групою Талібан, з точністю 84,74% — як злочин, вчинений групою Революційних збройних сил Колумбії, з точністю 79,71% — як злочин, вчинений Аль-Шабабом (Молодіжний рух Моджахедів), з точністю 76,39% — як злочин,

вчинений групою Боко Харам, з точністю 72,63% — як злочин, вчинений терористичним угрупуванням Донецька народна республіка.

Найбільший вплив на визначення терористичної групи має регіон, в якому стався терористичний акт (*region*), що вказує на те, що більшість терористичних груп діють на певних територіях і рідко за межами регіону (рис. 7).



Variable	Label	Number of Splitting Rules	Train Importance	Validation Importance
region		1	1,0000	1,0000
targsubtype1		5	0,5010	0,4568
nkill		3	0,1942	0,1630
weaptype1		1	0,1909	0,1994
targtype1		2	0,0652	0,0751
property		0	0,0000	0,0000
attacktype1		0	0,0000	0,0000
nwound		0	0,0000	0,0000
crit3		0	0,0000	0,0000
multiple		0	0,0000	0,0000
crit1		0	0,0000	0,0000
crit2		0	0,0000	0,0000
success		0	0,0000	0,0000
suicide		0	0,0000	0,0000

Рисунок 7 – Вплив змінних на класифікацію терористичних актів за терористичними угрупуваннями, які за них відповідають

Також значний вплив має тип цілі (*targsubtype1*), на який направлений терористичний акт. Тобто для класифікації є важливою ціль терористів: бізнес (комерційна діяльність організацій), уряд, поліція (як члени так і інститути), армія, аеропорти, дипломатичні установи, освітні заклади, журналісти і медіа, продукти харчування чи вода (склади, інфраструктура забезпечення), морські порти та бази, недержавні організації (Червоний хрест, Лікарі без кордонів). Високий вплив на результат також мають тип використаної зброї (*weaptype1*) та кількість вбитих (*nkill*), що підкреслює суспільну небезпеку тероризму.

Також в SAS Enterprise Miner реалізована можливість побудови автономного дерева рішень в автоматичному режимі. Змінивши у налаштуваннях модулю максимальну кількість гілок на 5, щоб перевірити залежність якості моделі від максимальної кількості гілок, на яку можна ділити один вузол, запускаємо блок автоматично, натиснувши у контекстному меню *Run*.

Здійснимо порівняльний аналіз побудованих моделей та вибір кращої з них за допомогою інструменту Порівняння моделей (блок *Model Comparison* на рис. 1).

Відбір кращої із побудованих моделей робився на основі мінімізації частки помилкової класифікації (*Misclassification Rate*, *MISC*) та такого показника, як середньоквадратична похибка (*Average Square Error*, *ASE*). Однак для прогнозів-рішень про якість моделі більше свідчить саме частка помилкової класифікації [7].

У табл. 2 наведено кількісні оцінки цих коефіцієнтів для кожної моделі по 2-х наборах даних. Можна побачити, що моделі характеризуються майже однаковими числовими характеристиками.

З табл. 2 видно, що інтерактивне дерево рішень є кращим за показником *Misclassification rate*, тоді як *Average square error* залишилась на попередньому рівні, що обґрунтовує доцільність вибору саме моделі інтерактивного дерева рішень.

Таблиця 2

Порівняння інтерактивного та автономного дерева рішень

№ п/п	Модель	Коефіцієнт помилкової класифікації (*Misclassification Rate*)		Середньоквадратична похибка (*Average Square Error*)	
		Навчальна	Валідаційна	Навчальна	Валідаційна
1	Інтерактивне дерево рішень	0,322	0,335	0,017	0,017
2	Автономне дерево рішень	0,414	0,436	0,016	0,017

Висновки з даного дослідження. В процесі розробки класифікаційної моделі визначення терористичної групи, яка відповідальна за скоєний терористичний акт було побудовано дерево рішень з використанням інструментів *SAS Enterprise Miner*. Визначено, що найвпливовішими факторами для класифікації є регіон, в якому вчинено терористичний акт, тип цілі, тип зброї та кількість загиблих в наслідок цього злочину.

Побудована модель може з точністю 94,12% класифікувати терористичний акт, як злочин, вчинений Республіканською армією Белуджистану, з точністю 92,55% — як злочин, вчинений групою Талібан, з точністю 84,74% — як злочин, вчинений групою Революційних збройних сил Колумбії, з точністю 79,71% — як злочин, вчинений Аль-Шабабом (Молодіжний рух Моджахедів), з точністю 76,39% — як злочин, вчинений групою Боко Харам, з точністю 72,63% — як злочин, вчинений терористичним угрупуванням Донецька народна республіка.

Подібні моделі можуть бути використані як національними так і міжнародними правоохоронними організаціями при проведенні попереднього аналізу терористичних угрупувань, що можуть бути причетні до певного терористичного акту з метою звуження кола підозрюваних.

Список літератури:

1. Шквірук В. Нові форми і методи тероризму в епоху глобалізації / В. Шквірук // Науковий вісник Чернівецького університету : Історія. Політичні науки. Міжнародні відносини. – 2013. (Вип. 676 – 677). – С. 225-229.
2. Яременко Н.С. Роль технологій Data Mining в боротьбі з тероризмом / Н.С. Яременко // Економіко-математичне моделювання: зб. мат. Першої нац. наук.-метод. конф., 30 вересня–1 жовтня 2016 р., м. Київ. – К. :КНЕУ, – 2016. – С. 402-403.
3. National Consortium for the Study of Terrorism and Responses to Terrorism START: A Center of Excellence of the U.S. Department of Homeland Security University of Maryland. Режим доступу : https://www.start.umd.edu/gtd/contact/.
4. SAS Enterprise Miner: деревья решений. Режим доступу : http://www.sas.com/content/dam/SAS/ru_ru/doc/academic/VMK_MGU/2015/lec4/EM04.pdf.
5. Богатырёв С. Введение в добычу данных (Data Mining) / С. Богатырёв, А. Симонова. – М. : БИНОМ, 2006. – 34 с. Режим доступа : http://yury.name/internet/01ia-seminar-note.pdf.
6. Інформаційні системи і технології в управлінні. Класифікація в бізнес-аналітиці. / Укл. : Біла Н.І. – Запоріжжя : ЗНТУ, 2014. – с. 50. Режим доступу до ресурсу: http://eir.zntu.edu.ua/bitstream/123456789/342/1/met_vk_bila_3.pdf.
7. Прикладная аналитика с использованием SAS Enterprise Miner 5: Материалы курса. Режим доступу: http://support.sas.com /software/products/miner/index.html.