Міністерство освіти і науки України

Сумський державний університет

Навчально-науковий інститут бізнес-технологій «УАБС»

Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

на тему «СТАТИСТИЧНО-ЙМОВІРНІСНА ОЦІНКА КОРУПЦІЙНИХ РИЗИКІВ В УКРАЇНІ»

Виконав студент 2 курсу, групи ЕК.м-81а

(номер курсу) (шифр групи)

Спеціальності 051«Економіка
(«Економічна кібернетика»)»

Дрозд С.А.

(прізвище, ініціали студента)

к.ф.-м.н., доцент Братушка С.М.

(посада, науковий ступінь, прізвище, ініціали)

Суми – 2019 рік

РЕФЕРАТ

кваліфікаційної магістерської роботи на тему

«СТАТИСТИЧНО-ЙМОВІРНІСНА ОЦІНКА КОРУПЦІЙНИХ РИЗИКІВ В УКРАЇНІ»

студента Дрозда Сергія Анатолійовича
(прізвище, ім’я, по батькові)

Актуальність теми даного дослідження обумовлюється необхідністю розробки дієвого механізму статистично-ймовірнісної оцінки корупційних ризикі.

Метою даної випускної кваліфікаційної роботи магістра є розробка моделі статистично-ймовірнісної оцінки корупційних ризиків в Україні.

Об’єктом даного дослідження є виступає корупційна злочинність в Україні.

Предметом дослідження даної роботи є методи та моделі побудови асоціативних правил та карт Кохонена.

Досягнення даної мети передбачає реалізацію наступних завдань Для досягнення поставленої мети необхідно вирішити наступні задачі:

– Дослідити поняття та особливості корупції.

– Проаналізувати рівень корупційних злочинів за останні роки в Україні та методи їх статистично-ймовірнісної оцінки.

– Визначити можливість використання методів Data Minig для статистично-ймовірнісної оцінки корупційних ризиків.

– Побудувати карти Кохенена та асоціативні правила для визначення закономірностей та статистично-ймовірнісної оцінки корупційних ризиків в Україні.

– Обґрунтувати вибір програмного забезпечення та провести розрахунки.

– Проаналізувати отримані результати.

– Визначити можливість практичного використання отриманих результатів.

Для досягнення поставленої мети та задач дослідження було проведено пошук асоціативних правил та побудова карт Кохонена.

Інформаційною базою кваліфікаційної магістерської роботи є публічна база корупційних порушень з даними з 2012 року по 3 квартал 2018.

Основний науковий результат кваліфікаційної магістерської роботи полягає у тому що було розроблено модель статистично-ймовірнісна оцінки корупційних ризиків в Україні яка включає в себе асоціативні правила та карти Кохонена.

Одержані результати дозволяють оцінювати ймовірність скоєння корупційних злочинів залежно від сфери діяльності та посади можливих злочинців, а також ймовірність того, що при скоєнні певного корупційного злочину додатково може бути скоєно інший корупційний злочин.

Ключові слова: асоціативні правила, корупція, карти Кохонена, правопорушення.

Зміст кваліфікаційної магістерської роботи викладено на 46 сторінках. Список використаних джерел із 50 найменувань, розміщений на 6 сторінках. Робота містить 1 таблицю, 21 рисунків, а також 4 додатків, розміщених на 8 сторінках.

Рік виконання кваліфікаційної роботи – 2019 рік.

Рік захисту роботи – 2019 рік.

Міністерство освіти і науки України
Сумський державний університет
Навчально-науковий інститут бізнес-технологій «УАБС»
Кафедра економічної кібернетики

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____
(науковий ступінь, вчене звання)

_____
(підпис) (ініціали, прізвище)
“___” _____20__ р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ МАГІСТЕРСЬКУ РОБОТУ
(спеціальність 051 «Економіка» («Економічна кібернетика»))
студенту 2 курсу, групи Ек.м.-81а
(номер курсу) (шифр групи)

Дрозд Сергій Анатолійович
(прізвище, ім’я, по батькові студента)

1. Тема роботи «Статистично-ймовірнісна оцінка корупційних ризиків в Україні»

затверджена наказом по університету від «____» _____ 2019 року №____

2. Термін подання студентом закінченої роботи «12» грудня 2019 року

3. Мета кваліфікаційної роботи: *розробка моделі статистично-ймовірнісної оцінки здійснення корупційних ризиків в Україні.*

4. Об’єкт дослідження: *корупційна злочинність в Україні.*

5. Предмет дослідження: *методи та моделі інтелектуальної обробки даних*

6. Кваліфікаційна робота виконується на матеріалах: *бази даних корупційних злочинів в Україні.*

7. Орієнтовний план кваліфікаційної роботи, терміни подання розділів керівникові та зміст завдань для виконання поставленої мети

Розділ 1 ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ СТАТИСТИЧНО-ЙМОВІРНІСНОЇ ОЦІНКИ КОРУПЦІЙНИХ РИЗИКІВ В УКРАЇНІ (до 11 листопада 2019 р.)

У розділі 1 необхідно дослідити предметну область, сутність поставленої задачі та проаналізувати існуючі методи статистично-ймовірнісні оцінки здійснення корупційних злочинів.

Розділ 2 РОЗРОБКА МОДЕЛІ СТАТИСТИЧНО-ЙМОВІРНІСНОЇ ОЦІНКИ КОРУПЦІЙНИХ РИЗИКІВ В УКРАЇНІ. (до 2 грудня 2019 р.)
У розділі 2 необхідно провести порівняльний аналіз методів інтелектуальної обробки даних, визначити методи аналізу, найбільш відповідні меті роботи, а також обгрунтувати вибір програмного забезпечення для проведення розрахунків.

Розділ 3 СТАТИСТИЧНО-ЙМОВІРНІСНА ОЦІНКА ЗДІЙСНЕННЯ КОРУПЦІЙНИХ ЗЛОЧИНІВ В УКРАЇНІ (до 9 грудня 2019 р.)
У розділі 3 провести налаштування даних, виконати розрахунки, необхідні для побудови карт Кохонена та визначення асоціативних правил з можливості здійснення корупційних правопорушень. Провести аналіз отриманих результатів.

8. Консультації з роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1			
2			
3			

9. Дата видачі завдання: «___»_____ 20__ року

Керівник кваліфікаційної роботи _____ Братушка С.М.
(підпис) (ініціали, прізвище)

Завдання до виконання одержав _____ Дрозд С.А.
(підпис) (ініціали, прізвище)

ЗМІСТ

ВСТУП

Проблема боротьби з корупцією актуальна сьогодні практично у всіх країнах світу. Дослідження західних економістів показують, що збільшення корупції на 1% гальмує економічне зростання країни на 0,4%, а скорочення корупційного тиску на 1% призводить до зростання ВВП на 0,72%.

В індексі сприйняття корупції (Corruption Perceptions Index, CPI), який складається міжнародною неурядовою організацією Transparency International, в 2019 році Україна займає 120-е місце з 32 балами. Прості розрахунки показують, що якби в найближчі п'ять років Україна могла вийти, наприклад, на рівень Чилі (27-й рейтинг, 67 балів), додаткові 35 балів сприйняття корупції могли б трансформуватися в економічне зростання на рівні понад 25%. Також доведено, що зростанню ВВП лише на 5% може дати Україні ефективні інструменти боротьби з бідністю і дозволить мінімізувати ризик трудової міграції [1].

Оскільки корупція в цілому, так само як і корупція при управлінні державними ресурсами, є складним системним явищем, то і дії по боротьбі з корупцією повинні носити системний характер. При цьому заходи, що пропонуються вченими і політиками, у вирішенні даної проблеми можна розділити на дві групи: каральні та превентивні. Останні спрямовані проти причин, а не зовнішніх виразів корупції. Боротьба з корупцією вимагає застосування рішень, в тому числі і спрямованих на передчасне виявлення закономірностей при корупційних злочинах. Це стає можливим в силу того, що найчастіше виявляється відразу кілька корупційних злочинів при порушенні кримінальної справи. Одним з інструментів виявлення названих закономірностей є методи DataMining. Саме це обумовлює актуальність тема роботи.

Метою роботи є розробка моделі статистично-ймовірнісної оцінки корупційних ризиків в Україні.

Об’єктом дослідження виступає корупційна злочинність в Україні.

Предмет роботи: методи та моделі побудови асоціативних правил та карт Кохонена.

Для досягнення поставленої мети необхідно вирішити наступні задачі:

– Дослідити поняття та особливості корупції.

– Проаналізувати рівень корупційних злочинів за останні роки в Україні та методи їх статистично-ймовірнісної оцінки.

– Визначити можливість використання методів Data Minig для статистично-ймовірнісної оцінки корупційних ризиків.

– Побудувати карти Кохенена та асоціативні правила для визначення закономірностей та статистично-ймовірнісної оцінки корупційних ризиків в Україні.

– Обґрунтувати вибір програмного забезпечення та провести розрахунки.

– Проаналізувати отримані результати.

– Визначити можливість практичного використання отриманих результатів.

РОЗДІЛ 1 ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ СТАТИСТИЧНО-ЙМОВІРНІСНОЇ ОЦІНКИ КОРУПЦІЙНИХ РИЗИКІВ В УКРАЇНІ

1.1 Поняття корупції в Україні та світі

Згідно з одним з визначень, корупцію можна розглядати як форму не чесності чи не законної діяльності, здійсненої особою чи організацією, на яку покладається посадова влада, часто для одержання незаконної користі або зловживання довіреною владою для особистої вигоди [2].

Незаконна діяльність може відбуватися в різних масштабах. Корупція варіюється від маленьких переваг між невеликою кількістю людей (дрібна корупція), до корупції, яка впливає на владу у величезному масштабі (велика корупція), і корупції, яка настільки розповсюджена, що є частиною повсякденної структури суспільства , включаючи корупцію як одну з ознак організованої злочинності. Корупція та злочинність - це соціологічні явища, які в різній мірі та пропорції часто зустрічаються практично у всіх державах світу. Кожні окремі держави виділяють внутрішні ресурси для нагляду та регулювання корупції та злочинності. Стратегії протистояння корупції часто узагальнюються терміном «антикорупція».

Науковець Стівен Д. Морріс пише, що політична корупція - це незаконне використання публічної влади для отримання приватних інтересів [3]. Економіст Ян Старший визначає корупцію як таємно надані дії, товар чи послугу третій стороні, щоб він чи вона могли впливати на певні види діяльності, які приносять користь корупціонеру, третій стороні або для всіх учасників зговору, у яких корумпований агент має повноваження [4]. Даніель Кауфман, співробітник Світового банку, поширює цю концепцію і на «юридичну корупцію», при якій влада зловживається в межах закону - оскільки особи, які мають владу, часто мають можливість приймати закони для захисту своїх незаконних інтересів [5]. Ефект корупції в інфраструктурі проявляться

у збільшенні затрат та часу будівництва, зниженні якості та зменшенні вигоди.

У нинішній час розроблено цілу низку показників та інструментів, які дозволяють вимірювати різні види корупції зі збільшенням точності.

Дрібна корупція зустрічається в менших масштабах і має місце під час виконання публічних послуг, наприклад, коли держслужбовці зустрічаються з суспільством [6, 7].

Велика корупція визначається як корупція, що виникає на найвищих рівнях влади таким чином, що потребує значного проникнення в політичну, правову та економічну систему [8]. Така корупція зазвичай зустрічається в державах з авторитарними чи диктаторськими урядами.

Урядова система багатьох держав ділиться на законодавчу, виконавчу та судову гілки, намагаючись надавати незалежні послуги, які менш піддаються великій корупції внаслідок незалежності гілок одна від одної [9].

Системна корупція - це корупція, яка пов'язана насамперед із слабкими сторонами організації чи процесу [10]. Це може протиставлятися окремим чиновникам чи агентам, які проводять незаконні дії в системі.

Фактори, які заохочують системну корупцію:

– включають суперечливі стимули;

– які не мають чітких меж дозволеного;

– монополістичні сили; відсутність прозорості;

– низька оплата виконаної праці;

– і культура безкарності [11].

Конкретні корупційні дії передбачають «хабарництво, вимагання та розтрату» в системі, де «корупція стає нормою, а не винятком» [12]. Науковці відрізняють централізовану та децентралізовану системну корупції, залежно від того, який рівень корупції у державі чи уряді спостерігається. Відзначимо, що в так званих пострадянських країнах, зустрічаються обидва типи корупції [13]. Деякі науковці стверджують, що існує обов'язок західних країн по захисту від систематичної корупції нерозвинених країн [14, 15].

Корупція досить довго була головним питанням у Азії, де суспільство сильно залежить від особистих стосунків до кінця XX століття, що поєдналося з новою спрагою до збагачення, призвело до збільшення рівня корупції. Історик Кіт Скоппа говорить, що хабарництво було лише одним із способів корупції в Китаї, який також включав «розтрачання, кумівство, обман, вимагання, відкати, шахрайство, контрабанду, розтрату грошових грошей, незаконні господарські операції, маніпуляції з запасами та шахрайство з нерухомістю». Враховуючи неодноразові антикорупційні кампанії, було доцільною обережністю відправити якомога більшу кількість шахрайських грошей за кордон [16].

Явище корупції може проявлятися у багатьох видах діяльності. Деякі з основних видів описані нижче.

Хабарництво це отримання уповноваженої особи фінансових або матеріальних цінностей від певної особи з метою одержання нечесної переваги в будь-якій діяльності такої особи [17].

Відкати є формою хабарництва, в якому комісія виплачуються хабарнику в обмін на надані послуги. Взагалі кажучи, про винагороду домовляються заздалегідь. Відкат відрізняється від інших видів хабарів тим, що мається на увазі договір між агентами двох сторін, а не з однією стороною, яка вимагає хабаря від іншої. Метою відкатів зазвичай є заохочення іншої сторони до співпраці в незаконній схемі [18].

Шахрайство передбачає використання фактів, що не відповідають дійсності, з метою переконати власника коштів або активів передати їх третій стороні [19].

Платежі по спрощенню. Це, як правило, невелика плата, щоб забезпечити або прискорити виконання завдання або необхідних дій, до яких платник має право, згідно з законодавством або іншим чином.

Благодійні та політичні внески, спонсорство, подорожі, і рекламні проекти є легальними видами діяльності для суб'єктів, але можуть бути зловживаннями і використовуватись в якості способів для хабарництва.

Конфлікт інтересів виникає, коли фізична або юридична особа, з обов'язком по збагаченню підприємства має конфлікт інтересів з особами з інших компаній, обов'язок або зобов'язання. Існування конфлікту інтересів само по собі звичайною справою, але корупція може виникнути тоді, коли управління компанії або співробітник за контрактом третьої сторони буде діяти в інтересі іншої [20].

Змова може приймати різні види, найбільш поширеними в цьому розумінні є картелі і фіксування цін [21].

Картель - це таємне угода або змова між підприємствами для здійснення корупційних дій або шахрайства. Як правило, це включає в себе встановлення цін, обмін інформацією, встановлення квоти на виробництво і поставку товарів [22].

Фіксування ціни - це угода між конкурентами по підняттю, виправленню чи іншому встановленню ціни, розповсюдження товарів або послуг. В такій ситуації потрібно, щоб учасники змови встановлювали точно таку ж ціну, а саме встановлення цін може приймати різні форми. Тому будь-яка угода, яка обмежує цінову політику компаній в умовах конкуренції може привести до порушення чинного законодавства в сфері конкуренції.

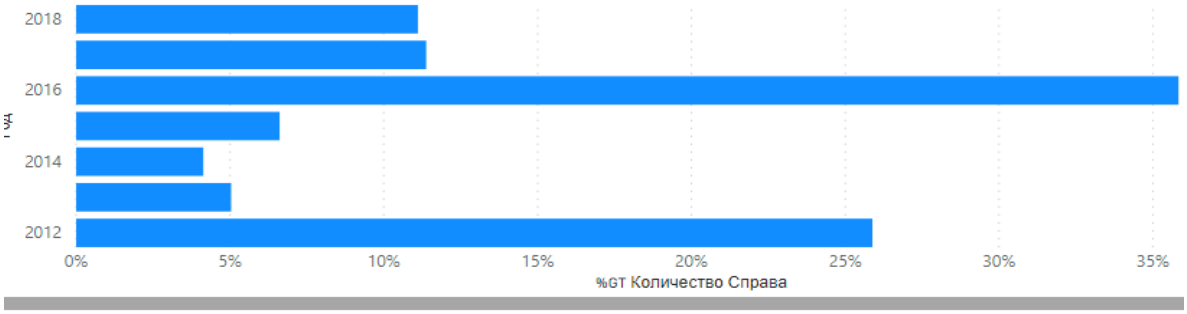
«Двері, що обертаються». Це корупційна схема, пов'язана з переміщенням з робочих місць працівників високого рівня, що працюють в державному секторі, на роботу в приватному секторі, і навпаки [23].

Фаворитизм передбачає корупцію, в якому обрана людина, незалежно від кваліфікації, заслуг, або прав, отримує привілеї через свої зв’язки [24].

Інсайдерська торгівля. Це торгівля непублічною інформацією компанії людиною, яка в ній працює [25].

На рисунку 1.1. приведено статистичні дані (по рокам) по корупційним злочинам в Україні за період з 2012 по 3-й квартал 2018 р. За аналізом рисунку можна зробити висновок, що на 2018 і 2017 роки припадає приблизно по 11 % від загальної кількості злочинів, в той час, як на 2016 рік майже 36 %, що можна пояснити початком роботи Національного антикорупційного бюро України (НАБУ) в 2016 р. В той же час на 2015, 2014 2013 роки припадає

відповідно 6,6%, 4,1 %, 5 % від загальної кількості (понад 25 тис. злочинів). Високий відсоток корупційної злочинності у 2012 році (близько 26 %) може бути пояснений сплеском активності органів правопорядку при зміні екс-президента В. Януковича.



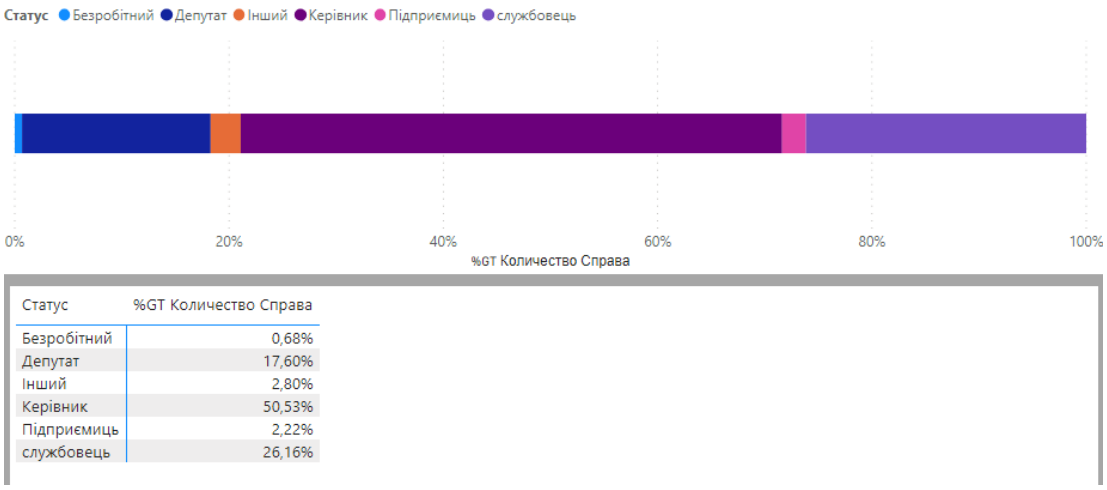
Год	%GT Количество Справа
2012	25,88%
2013	5,04%
2014	4,13%
2015	6,62%
2016	35,84%
2017	11,38%
2018	11,11%

Рисунок 1.1 – Розподіл даних по рокам

Досить цікаву інформацію можна отримати шляхом аналізу кількості корупційних правопорушень в розрізі посад, на яких було здійснено корупційні злочини (рис. 1.2, рис. 1.3).

Год	Безробітний	Депутат	Інший	Керівник	Підприємиць	службовець
2012	0,17%	0,53%	1,00%	15,36%	0,98%	7,85%
2013		0,39%	0,23%	3,17%	0,05%	1,20%
2014	0,01%	0,88%	0,10%	2,24%	0,03%	0,87%
2015	0,04%	1,46%	0,15%	3,54%	0,11%	1,33%
2016	0,37%	6,75%	0,61%	18,12%	0,87%	9,11%
2017	0,08%	3,51%	0,29%	4,46%	0,11%	2,93%
2018	0,01%	4,08%	0,44%	3,65%	0,06%	2,87%

Рисунок 1.2 – Розподіл кількості корупційних злочинів в Україні по рокам в розрізі посад



Статус	%GT Количество Справа
Безробітний	0,68%
Депутат	17,60%
Інший	2,80%
Керівник	50,53%
Підприємиць	2,22%
службовець	26,16%

Рисунок 1.3 – Розподіл кількості корупційних злочинів в Україні за всі роки спостережень в розрізі посад

З приведених рисунків можна зробити висновок, побачити, до корупційних ризиків найбільш схильні люди, що працюють на керівних посадах, наданою їм владою (50,53% від загальної кількості), держслужбовці (26,16%) і працівники органів місцевого самоврядування (17,60%).

З кругової діаграми, представленої на рисунку 1.4, ми бачимо:

- найбільшу кількість корупційних правопорушень було здійснено за ст. 172 – «Грубе порушення законодавства про працю» (11951 кримінальних справ або 49.13% від загальної кількості);

ст. 191 – «Привласнення, розтрата майна або заволодіння ним шляхом зловживання службовим становищем» (3548 та 14.59% відповідно);

ст. 368 – «Прийняття пропозиції, обіцянки або одержання неправомірної вигоди службовою особою» (2827 та 11.62%);

ст. 366 – «Службове підроблення» (2024 та 8.32%);

ст. 364 – «Зловживання владою або службовим становищем» (1546 та 6.36%);

ст. 369 – «Пропозиція, обіцянка або надання неправомірної вигоди службовій особі» (1170 та 4,81%). [26].

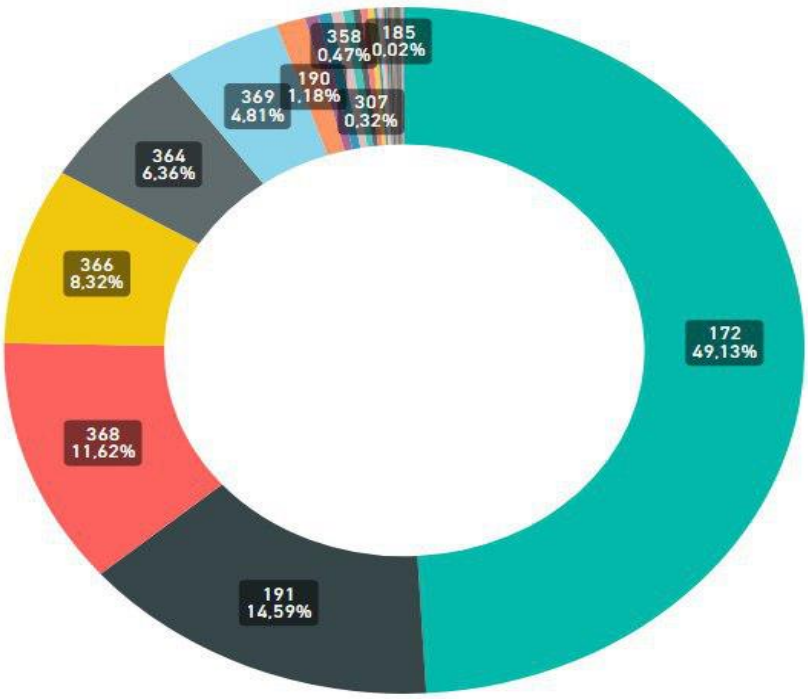


Рисунок 1.4 – Розподіл правопорушень корупційних злочинів за статтями кримінального кодексу

1.2.Методи статистично-ймовірнісної оцінка корупційних ризиків

Ймовірнісно-статистичні методи аналізу передбачать як оцінку ймовірності виникнення події, так і розрахунок відносних ймовірностей того чи іншого варіанту розвитку подій в майбутньому. При цьому аналізуються розгалужені ланцюжки подій і наслідків, вибирається відповідний математичний апарат і оцінюється ймовірність реалізації події. Основні обмеження такого підходу пов’язані з недостатньою кількістю статистичних даних. Окрім того, використання спрощених розрахункових схем знижує достовірність отриманих оцінок ризику. Тим не менш, ймовірнісний метод в даний час вважаться одним з найбільш поширених. На його основі розроблено різні методики оцінки реалізації подій, які відносно від вхідної інформації поділяться на:

– *ймовірнісно-статистичні*, коли ймовірності визначаються за доступними статистичними даними;

– *теоретико-ймовірнісні* - застосовуються для оцінки ризиків від нечастих подій, чи коли статистика практично відсутня;

– *ймовірнісно-евристичні*, засновані на застосуванні суб'єктивних ймовірностей, отриманих за допомогою експертного оцінювання. Такі методики використовуються при оцінці комплексних ризиків від сукупності небезпек, коли відсутні не тільки статистичні дані, але і математичні моделі, або точність математичних моделей занадто низька.

До ймовірнісно-статистичних методів відносяться:

– статистичні *якісн*і методи: карти потоків;

– *кількісні* методи: контрольні карти.

До теоретико-ймовірнісних методів відносять:

– *якісні*: причини послідовності;

– *кількісні*: аналіз дерев подій; аналіз дерев відмов; оцінка ризику мінімальних маршрутів від ініціюючої до основної події; дерево рішень; ймовірнісна оцінка ризику.

До ймовірнісно-евристичних методів належать:

– *якісні*: метод експертного оцінювання, метод аналогій;

– *кількісні*: метод банальних оцінок, метод суб'єктивних ймовірностей оцінки небезпечних станів, метод узгодження групових оцінок.

Ймовірнісно-евристичні методи застосовуються при малій кількості статистичних даних і у випадку рідкісних подій, коли можливості використання точних математичних методів обмежені через відсутність мінімально потрібних інформаційних даних про показники надійності і параметри досліджуваних об'єктів, а також через відсутність перевірених математичних моделей, що посилаються на реальний стан системи. Ймовірнісно-евристичні методи ґрунтуються на застосуванні суб'єктивних ймовірностей, що розраховуються методами експертного оцінювання [27].

Імовірність кожної ідентифікованої не законної схеми слід оцінювати без урахування контролю на місці і на підприємстві. Іншими словами підприємство, де можливості для здійснення корупційного злочину обумовлені відсутністю достатнього контролю зовні. Окремо слід розглянути можливості корупційних злочинів, які здійснюються індивідуально або групою осіб.

В рамках такої структури, рекомендується, щоб оцінка ймовірності була розрахована як ймовірність корупційних дій, що відбуваються протягом наступних 12 місяців. Такі часові рамки повинні бути скореговані в міру необхідності, щоб відповідати характеристикам цілей управління ризиком корупційного злочину.

Відзначимо, що в оцінці потенційної можливості скоєння корупційного злочину деякі фахівці вважають за доцільне визначати певні категорії (наприклад, сферу діяльності, посаду чи інші показники) і визначати ймовірність здійснення корупційних злочинів за певними статтями залежно від конкретних значень (посад, галузей тощо).

Якщо обчислювальні ресурси достатні, можна віддати перевагу підходу, коли береться до уваги відразу декілька критеріїв, наприклад пара «посада-галузь», або пара «стаття-посада» [28].

РОЗДІЛ 2 РОЗРОБКА МОДЕЛІ СТАТИСТИЧНО-ЙМОВІРНІСНОЇ ОЦІНКИ КОРУПЦІЙНИХ РИЗИКІВ В УКРАЇНІ

2.1 Методи дослідження

DataMining - це процес знаходження шаблонів (закономірностей) у великих наборах даних із залученням методів машинного навчання, статистики та систем баз даних [29]. DataMining - це міждисциплінарна основа інформатики та статистики, що має загальну мету - отримання нових знань інтелектуальними методами з набору вже відомих даних та перетворення отриманої інформації в зрозумілу структуру даних для подальшого використання в дослідженнях та практичній діяльності [30].

Обробка даних – це один з етапів аналізу процесу «виявлення знань у базах даних» або KDD [31]. Окрім покрокового аналізу, він також включає аспекти управління базами даних та даних, попередню обробку даних, міркування щодо моделей та висновків, пост-обробку виявлених структур, візуалізацію та оновлення. Метою даного етапу є вилучення відомих шаблонів і знань з великої кількості даних, а не отримання самих даних [32]. На сьогодні цей термін є досить поширеним і у широкому розумінні мається на увазі будь-яка форма широкомасштабної обробки даних або інформації (збирання, вилучення, складування, аналіз та статистика), а також будь-яке застосування комп'ютерної системи підтримки прийняття рішень, включаючи штучний інтелект (наприклад, машинне навчання) та бізнес-інтелект [33]. Часто більш загальні терміни, такі як «масштабний аналіз даних та аналітика» або, якщо йдеться про фактичні методи, «штучний інтелект та машинне навчання» - є більш коректними.

Актуальним завданням пошуку нових знань є напівавтоматичний або автоматичний аналіз великої кількості даних для вилучення раніше невідомих, цікавих зразків, таких як групи записів даних (кластерний аналіз), не-

звичайні записи (виявлення аномалії) та залежності (отримання асоціативних правил). Зазвичай це передбачає використання методів баз даних, таких як просторові індекси. Важливим є те, що ні збір та підготовка даних, ні інтерпретація результатів та звітування не є частинами отримання даних, але вони можуть бути віднесені до загального процесу KDD як додаткові кроки.

Різниця між аналізом даних та обробкою даних полягає в тому, що аналіз даних використовується для тестування моделей та гіпотез набору даних, наприклад, аналізу ефективності маркетингової кампанії незалежно від кількості даних. На противагу цьому, для обміну даними використовується машинне навчання та статистичні моделі для розкриття нових прихованих правил (залежностей) у великих обсягах даних [34].

Пов'язаний з цим процесом термін Data dredging (отримання даних) передбачає використання методів отримання даних для вибірки частин загального набору даних, які можуть бути занадто малі, щоб зробити достовірні статистичні висновки про достовірність виявлених закономірностей. Ці методи, однак, можуть бути використані при створенні нових гіпотез для перевірки на сукупності відомих даних.

2.2. Асоціативні правила

Навчання правилам асоціацій - це метод машинного навчання на основі правил для виявлення цікавих зв’язків між змінними у великих базах даних. Він призначений для встановлення чітких правил, виявлених в базах даних, використовуючи певні методи асоціацій.

Спираючись на концепцію чітких правил, Ракеш Аграваль, Томаш Імелієнський та Арун Свамі запровадили правила асоціації для виявлення закономірностей між продуктами, що одночасно придбаються, при аналізі даних про транзакції, зафіксовані системами торгових точок (POS) у супермаркетах [35]. Така інформація може бути використана як основа для прийн-

яття рішень щодо маркетингової діяльності, таких як, наприклад, рекламні акції або оптимального розміщення товарів.

На додаток до вищенаведеного прикладу застосування асоціативних правил для аналізу ринкових кошиків такі ж підходи використовуються у багатьох сферах діяльності людини, включаючи розробку веб-додатків, виявлення вторгнень, безперервне виробництво та біоінформатику. На відміну від виявлення послідовностей подій, у процесі побудови асоціативних правил зазвичай не враховується порядок позицій ні в межах транзакції, ні через транзакції. Асоціативні правила лише допомагають виявити зв’язки між елементами у великих базах даних. Ці правила не встановлюють переваги окремих осіб, скоріше знаходять зв’язки між набором елементів кожної окремої транзакції. Саме це відрізняє їх від спільної фільтрації.

Для встановлення залежностей у правилах транзакції не обмежують протягом часу. Список елементів з унікальними ідентифікаторами транзакцій вивчається як одна група. З іншого боку, спільна фільтрація зв'язує всі транзакції, що відповідають ідентифікатору користувача, щоб виявити схожість між налаштуваннями користувачів. Це корисно для рекомендування елементів користувачам, наприклад, на веб-сайтах електронної комерції.

Розглянемо докладніше, як виглядає асоціативне правило. Воно складається з причини та наслідку, які у свою чергу є переліком пунктів. Зауважимо, що причиною є спільне виникнення, а не наслідок. Для даного правила набір елементів - це список усіх елементів у попередньому та наступному.

Існують різні показники (характеристики) асоціативних правил, які допомагають зрозуміти силу зв'язку між змінними.

1. Підтримка. Цей показник дає уявлення про частоту (кількість) набору предметів у всіх транзакціях. Математично підтримка - це частка від загальної кількості транзакцій, в яких відбувається набір елементів.

$$\text{Support}(\{\text{X}\} \rightarrow \{\text{Y}\}) = \frac{X \cup Y}{ID}, \tag{2.1}$$

де X і Y - елементи асоціативного правила;

ID загальна кількість транзакцій.

Значення підтримки допомагає нам визначити правила, які варто враховувати для подальшого аналізу.

2. Довіра. Цей показник визначає ймовірність появи наслідків у наборі даних, враховуючи, що для таких наборів вже є прецеденти. Технічно довіра до правила - це умовна ймовірність виникнення наслідків з урахуванням попередніх:

$$Conf(\{X\} \longrightarrow \{Y\}) = \frac{ID(X \cup Y)}{ID(X)} \tag{2.2}$$

Довіра до асоціативного правила з наслідками, що часто повторюються, завжди буде високою.

3. Ліфт - це відношення довіри до базової ймовірності виникнення {Y}

$$Lift(\{X\} \longrightarrow \{Y\}) = \frac{\frac{ID(X \cup Y)}{ID(X)}}{ID(Y)} \tag{2.3}$$

У випадках, коли {X} насправді веде до {Y} у наборі, значення ліфту буде більше 1.

У загальному випадку асоціативні правила будуються за алгоритмом Apriori [36].

1. Генерування наборів елементів (транзакцій) із списку елементів

Першим кроком у створенні правил асоціації є отримання всіх частих наборів елементів, на яких можна виконати бінарні розділи, щоб отримати попередній та наступний.

Часті набори предметів - це ті, які відбуваються принаймні мінімальну кількість разів в операціях (транзакціях). Технічно це набори елементів, для яких значення підтримки (частка транзакцій, що містять набір елементів) перевищує мінімальний поріг, тобто проміжок часу.

Підхід повного перебору (грубої сили) для пошуку частих (популярних) наборів подій полягає у формуванні всіх можливих наборів елементів та перевірці значення підтримки кожного з них. Алгоритм Apriori допомагає зробити такий пошук більш ефективним. У ньому йдеться про те, що є результатом немонотонності показника підтримки. Принцип Apriori дозволяє ігнорувати всі набори елементів, які не відповідають умові мінімального порогового рівня для підтримки.

Методологія, яка дозволяє отримати результат, називається алгоритмом Apriori. Необхідні кроки:

1. Формуються всі часті набори елементів (підтримка ≥ min_sup), що містять лише один елемент. Далі генеруються набори елементів довжиною 2 як усі можливі комбінації вищевказаних наборів елементів. Потім ігноруються ті набори, для яких значення підтримки є нижче мінімального. Далі генеруються набори елементів довжиною 3, як усі можливі комбінації наборів довжини 2 (які залишилися після скорочення) та здійснюється така сама перевірка значення рівня підтримки. У подальшому продовжують збільшувати довжину наборів елементів на один і перевіряється відповідність пороговому значенню на кожному кроці.

2. Генеруються всі можливі правила із популярних (частих) наборів елементів. Правила формуються бінарним розділом кожного набору елементів.

Із переліку всіх можливих правил мається на меті визначити правила, які опускаються вище мінімального рівня довіри (min_conf). Так само, як використовується немонотонність підтримки, довіру у правилах, створених з одного набору елементів, також слід перевіряти на немонотонність. Це застосовується відносно кількості елементів, що є наслідком.

Це означає, що довіра у (A, B, C → D) ≥ (B, C → A, D) ≥ (C → A, B, D). Нагадаємо, характеристика «довіра» асоціативного правила визначається співвідношенням (2.2).

Як ми знаємо, підтримка всіх правил, створених з одного набору елементів, залишається однаковою, і різниця виникає лише в розрахунку значення показника довіри у знаменнику (2.2). Зі зменшенням кількості елемен-

тів у X збільшується підтримка {X} (що випливає з антимонотонності властивості підтримки), а отже, і значення довіри зменшується [37].

На практиці значення для параметрів мінімальна (максимальна) підтримка і мінімальна (максимальна) достовірність вибираються таким чином, щоб обмежити кількість знайдених правил. Якщо підтримка має велике значення, то алгоритми будуть знаходити правила, добре відомі аналітикам або настільки очевидні, що немає ніякого сенсу проводити такий аналіз.

З іншого боку, низьке значення підтримки веде до генерації значної кількості правил, що, зазвичай, вимагає суттєвих обчислювальних ресурсів. Більшість нових правил знаходиться саме при низькому значенні порогу підтримки, хоча занадто низьке значення підтримки правила веде до генерації статистично обґрунтованих правил.

Асоціативні правила з високою підтримкою можуть застосовуватися для формалізації добре відомих правил, наприклад, в автоматизованих системах для управління процесами або персоналом.

Треба відзначити, що поняття «висока» і «низька» підтримка або достовірність дуже сильно залежать від предметної області.

Концептуальна модель побудови асоціативних правил стосовно теми обраного дослідження може бути подана наступним чином (рис. 2.1):

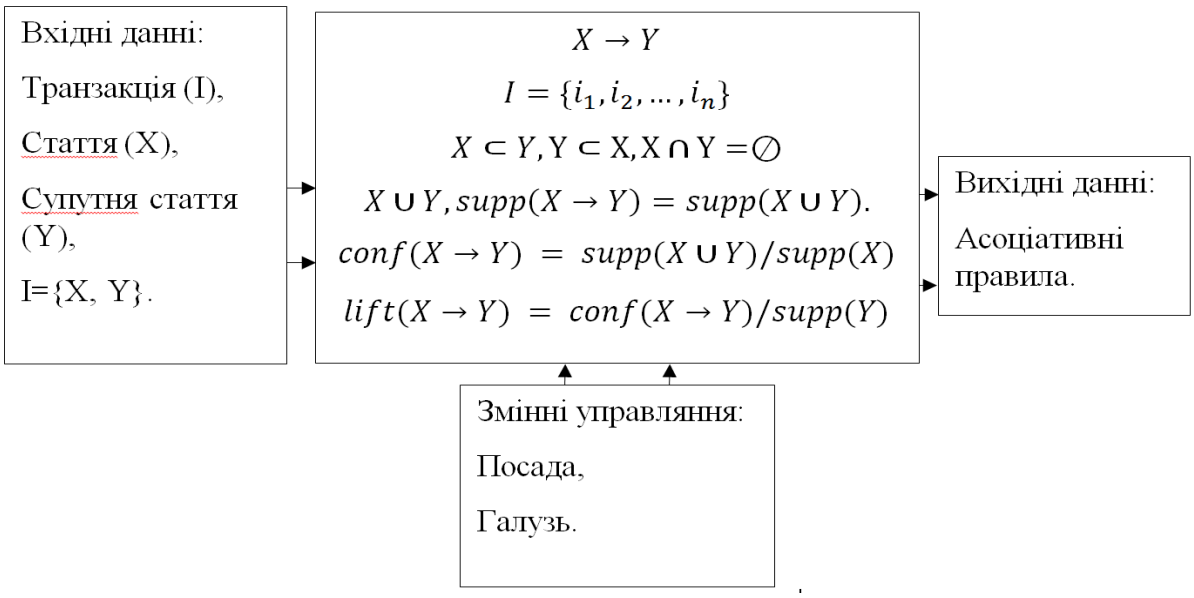


Рисунок 2.1 – Концептуальна модель пошуку асоціативних правил

2.3. Карти Кохонена

Карта, що самоорганізовується (SOM) - це тип штучної нейронної мережі (ANN), яка навчається за допомогою некерованого навчання для створення n-мірного (як правило, двовимірного) дискретного подання вхідного простору навчальних зразків, званого картою, і є методом зменшення розмірності початкового багатовимірного простору даних. Карти, що самоорганізовуються, відрізняються від інших штучних нейронних мереж, оскільки в основу їх побудови покладено принцип конкурентного навчання, на відміну від навчання виправлення помилок (наприклад, зворотне розповсюдження з градієнтним спуском), і в такому сенсі, при побудові карт використовують функцію сусідства для збереження топологічних властивостей вхідного простору.

Принцип SOM було запропоновано фінським професором Тьюво Кохоненом у 1980-х роках, і тому його іноді називають картою Кохонена [38].

Карта, що самоорганізовується (карти Кохонена) можуть використовуватися для вирішення задач наступних типів: моделювання, прогнозування, пошук закономірностей у великих масивах даних, виявлення наборів незалежних ознак і стиснення інформації.

Алгоритм функціювання самоорганізованих карт (Self Organizing Maps-som) являє собою один з варіантів кластеризації багатовимірних векторів з збереженням топологічної подоби.

Прикладом таких алгоритмів може бути алгоритм k-найближчих середніх (k-means). Важливою відмінністю алгоритма SOM це те, що всі нейрони (вузли, центри класів) впорядковані в деяку структуру (двовимірну сітку). При цьому в ході навчання модифікується не тільки нейрон-переможець, але і його сусіди, хоча і в менший мірі. За рахунок цього SOM можна вважати як один з методів проектування багатовимірного простору в простір з більш низькою розмірністю. При використанні цього алгоритму вектори, близькі у вихідному просторі, виявляються поруч і на отриманій карта.

Алгоритм SOM передбачає використання впорядкованої структури нейронів. Зазвичай використовуються одно- і двовимірні сітки. При цьому кожен нейрон являє собою n-мірний вектор-стовбець, де n визначається розмірністю вхідного простору значень (розмірністю вхідного вектору). Застосування одно - і двовимірних сіток пов'язано з тим, що виникають проблеми при відображення просторових структур більшої розмірності (при цьому виникають проблеми із зниженням розмірності до двовимірної - представлення на моніторі).

Зазвичай нейрони розташовуються у вузлах двовимірної сітки з прямокутними або шестикутними осередками. При цьому, як було описано вище, нейрони також взаємодіють один з одним. Величина такої взаємодії визначається відстанню між нейронами на карті.

Як правило, при реалізації алгоритму SOM заздалегідь задається конфігурація сітки (прямокутна або шестикутна), а також кількість нейронів в мережі. Деякі літературні джерела рекомендують використовувати максимально можливу кількість нейронів в карті. При цьому початковий радіус навчання значною мірою впливає на здатність узагальнення за допомогою отриманої карти. У разі, коли кількість вузлів карти перевищує кількість прикладів у навчальній вибірці, успіх використання алгоритму великою мірою залежить від відповідного вибору початкового радіусу навчання. Однак, в випадку, коли розмір карти становить десятки тисяч нейронів, час, необхідний для навчання карти, зазвичай буває занадто великим для вирішення практичних завдань. Таким чином, необхідно дотримуватись допустимого компромісу при виборі кількості вузлів.

Алгоритм побудови карт Кохонена наступний:

Використані позначення:

t - номер ітерації,

n - розмірність початкового простору,

$x \subset R^n$ - навчальна вибірка,

ф $\subset R^n$ - тестова вибірка,

$N = |x|$ - розмір навчальної вибірки,

$K = |ф|$ - розмір тестової вибірки,

α - висота шару Кохонена,

β - ширина шару Кохонена,

$M = \alpha \times \beta$ - загальна кількість нейронів в шарі Кохонена,

$ne_j, j = 1, \dots M$ - нейрон з індексом j

$$x^i(t) = \left(x_1^{(i)}(t), \dots, x_n^{(i)}(t)\right), i = 1, \dots, N \quad (2.4)$$

де i- номер вектору, призначеного для кластеризації,

$$w^{(j)}(t) = \left(w_1^{(j)}(t), \dots, w_n^{(j)}(t)\right), j = 1, \dots, M \quad (2.5)$$

де j- номер вектору вагових коефіцієнтів для j -го нейрона,

r_k - положення k -го нейрона в топології карти, тобто (i_k, j_k),

$h_{ij}(t)$ - функція сусідства між нейронами з індексами i і j на ітерації t,

$h(d, t)$ - функція від відстані між нейронами,

$\alpha(t)$ - швидкість навчання мережі,

$\sigma(t)$ - функція, яка зменшує кількість сусідів з при збільшенні ітерації (монотонно спадною),

T - максимальне число ітерацій алгоритму навчання - може бути задано фіксоване, або ж дорівнює потужності навчальної вибірки, тобто

Тоді.

Крок №0. Визначення значень параметрів алгоритму: використовувані метрики, $\alpha(t), \sigma(t) t \coloneqq 0$ функції

Крок №1. Ініціалізація векторів початкових ваг нейронів

$$w^{(j)}(0) \forall j = 1, \dots, M \quad (2.6)$$

Крок №2. Якщо в навчальній вибірці ще є елементи, отримати випадковим чином (наприклад, індекс i - може бути реалізацією випадкової величини з рівномірного дискретного розподілу на $[0, N]$) вхідний вектор,

$$x^{(i)}(t), N \coloneqq N - 1, x = x \backslash x^{(j)}(t) \tag{2.7}$$

В іншому випадку алгоритм завершує свою роботу

Крок №3. Для

$$x^{(j)} \text{ і } w^{(j)}(t) \forall j = 1, \dots, M \tag{2.8}$$

Знайти

$$p\left(x^{(i)}(t) - w^{(j)}(t)\right) = \left\|x^{(i)}(t) - w^{(j)}(y)\right\|^2 \tag{2.9}$$

Крок №4. Знаходимо нейрон-переможець

$$n\epsilon_c, \text{c} = arg_{\forall j \in \{1,\dots M\}} \min p(x^{(i)}(t) w^{(i)}(t)), \tag{2.10}$$

який лежить ближче до поточного об'єкту $x^{(i)}(t)$ по даній метриці $\|°\|$

Крок №5. Для всіх нейронів $n\epsilon_j$ виконати перерахунок вагових векторів:

$$w^{(i)}(t+1) \coloneqq w^{(j)}(t) + \alpha(t) h_{ci}(t) \left[x^{(i)}(t) - w^{(j)}(t)\right] \tag{2.11}$$

Крок №6. $t \coloneqq t + 1$

Крок №7. Перейти до кроку 2.

На цьому алгоритм побудови карт Кохонена закінчено [39].

Недоліки використання карт Кохонена:

– модель карт не будує генеративної моделі для даних, тобто модель не розуміє, як створюються дані;

– модель гарно працює при дослідженні категоріальних даних, але набагато гірше для даних змішаних типів;

– час, необхідний для підготовки моделі, досить значний, а сама мережа важко тренується при дослідженні даних, що незначно змінюються [40].

Деякі особливості підготовки навчальної вибірки.

– Основна відмінність у підготовці навчальної вибірки для навчання нейромережі в алгоритмі полягає в тому, що вихідні поля в такій вибірці можуть бути відсутні зовсім. Навіть якщо в навчальній вибірці будуть присутні вихідні поля, вони не братимуть участі при навчанні нейромережі, однак при цьому вони все рівно будуть брати участь при відображенні карт.

– Нормалізація полів така ж, як для звичайних нейромереж. Налаштування навчальної вибірки Навчальна вибірка налаштовується також, як і для нейромережі і дерева рішень.

– Перед початком навчання карти необхідно про ініціалізувати вагові коефіцієнти нейрон. Вдало обраний спосіб ініціалізації може істотно прискорити навчання і привести до отримання більш якісних результатів.

Існує три способи ініціалізації початкових ваг.

– Випадковими значеннями, коли всім вагам надаються малі випадкові величини.

– З навчального множини, коли в якості початкових значень задається значеннями випадково вибраних прикладів з навчальної вибірки;

– З власних векторів. У цьому випадку ваги ініціюються значеннями векторів, лінійно впорядкованих вздовж лінійного підпростору, що проходить між двома головними власними векторами вихідного набору даних.

Отриману в результаті навчання карту можна представити у вигляді листкового пирога, кожен шар якого являє собою розмальовку, породжену однією з компонент вихідних даний. Такий набір розмальовок може використовуватись для аналізу закономірностей, що існують між компонентами набору даних. Після формування карти отримують набір вузлів, який можна відобразити у вигляді двовимірної картинки.

При цьому кожному вузлу карти можна поставити у відповідність ділянку на рисунку (чотири- або шестикутний), координати якого визначаються координатами відповідного вузла в решітці. Для візуалізації результатів залишається лише визначити колір осередків. Для цього і використовуються значення компонент.

Найпростіший варіант – використання градацій сірого. В цьому випадку комірки, відповідні вузлам карти, в які потрапили елементи з мінімальними значеннями компонента або не потрапило взагалі жодного запису, будуть зображені чорним кольором, а комірки, в які потрапили записи з максимальними значеннями такого компонента, будуть відповідати осередку білого кольору. Більш зручною є використання для розмальовки кольоровий палітри. В принципі можна використовувати будь-яку градієнтну палітру для розмальовки. Приклад створених карт Кохонена приведено на рис. 2.2.

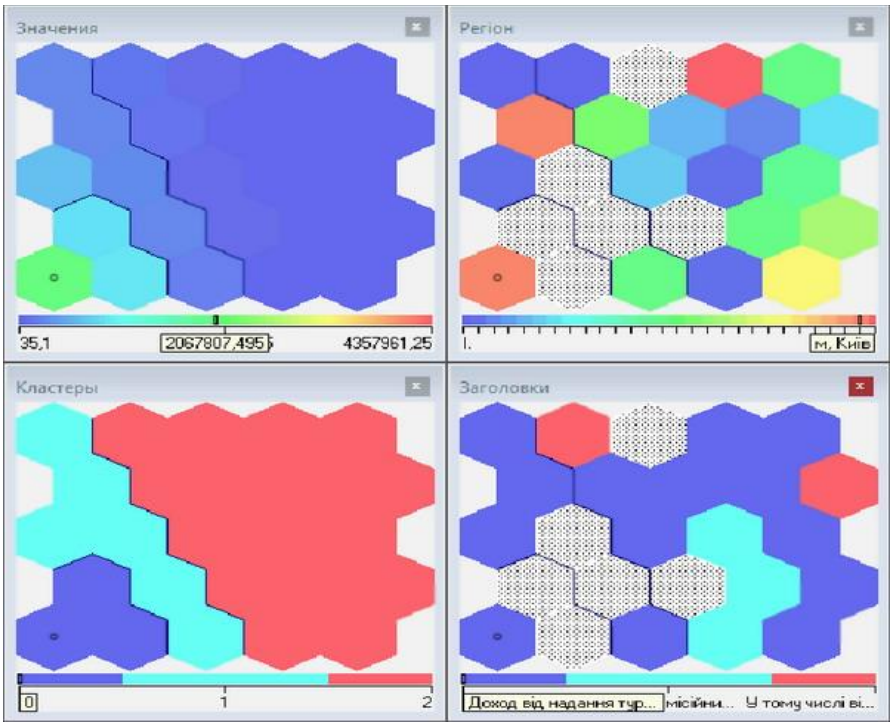


Рисунок 2.2 – Варіант побудованих карт Кохонена

Отримані розмальовки в сукупності утворюють атлас, що відображає розташування компонент, зв'язку між ними, а також відносне розташування різних значень компонент [41].

2.4 ПЗ з реалізацією методів статистично-ймовірнісного оцювання

На сьогодні для проведення наукових досліджень та прогнозування використовується значна кількість програмних пакетів спеціалізованого та універсального призначення. Наведемо перелік найбільш поширених з них.

Statistica - програмний пакет для аналітики, який забезпечує аналіз даних (зокрема кластерізацію та факторний аналіз), управління даними, визначення статистичних характеристик, обмін даними та має широкі можливості з візуалізації результатів [42]. Окремо відзначимо, що даний пакет додатково дозволяє створювати та використовувати нейронні мережі для певних класів задач.

Orange - візуалізація даних із відкритим кодом, машинне навчання та інструментарій DataMining. Він дозволяє виконувати візуальне програмування переднього плану для дослідницького аналізу даних та інтерактивної візуалізації даних . Orange допомагає сформувати дані та є програмним забезпеченням на основі компонентів. Він написаний обчислювальною мовою Python [43].

Rapid Miner - одна з систем прогнозування та аналізу, розроблена компанією з однойменною назвою Rapid Miner. Програмний засіб написано мовою програмування JAVA. Він забезпечує інтегроване середовище для глибокого навчання, дослідження тексту, машинного навчання та прогнозного аналізу. Інструмент може бути використаний для широкого спектру застосувань, включаючи бізнес-програми, комерційні програми, навчання, освіту, дослідження, розробку програм, машинне навчання. В якості основи системи знаходиться модель «клієнт/сервер». Rapid Miner поставляється з шаблонами, які дають можливість швидкої доставки даних із зменшеною кількістю помилок (що зазвичай очікується в процесі написання ручного коду). Rapid Miner складається з трьох модулів:

Rapid Miner Studio - модуль призначено для проектування, складання прототипів, валідації тощо;

Rapid Miner Server - для управління прогнозними моделями даних, створеними в студії;

Rapid Miner Radoop - виконує процеси безпосередньо в кластері Hadoop для спрощення прогнозного аналізу [44].

Waikato Environment - програмне забезпечення машинного навчання, розроблене в університеті Вайкато в Новій Зеландії. Він підходить для аналізу даних та прогнозного моделювання. Він містить алгоритми та засоби візуалізації, які підтримують машинне навчання. У Weka є графічний інтерфейс, який полегшує доступ до всіх функцій. Він написаний мовою програмування JAVA. Weka підтримує основні завдання з пошуку даних, включаючи обробку даних, обробку даних, візуалізацію, регресію тощо. Це працює за умови, що дані доступні у вигляді плоского файлу. Weka може надати доступ до баз даних SQL через підключення до бази даних та може додатково обробляти дані чи результати, повернуті запитом [45].

KNIME - інтеграційна платформа для аналізу даних та звітності, розроблена KNIME.com.ag. ПЗ працює за концепцією модульного конвеєра даних. KNIME складається з різних компонентів машинного навчання та обміну даними, вбудованих разом. KNIME широко застосовується для фармацевтичних досліджень. Окрім того, він чудово працює для аналізу даних клієнтів, аналізу фінансових даних та бізнес-аналізу. [46].

Sisense є надзвичайно корисним і найкращим чином підходить для програмного забезпечення BI, що стосується цілей формування звітності в організації. Він має можливість обробляти та обробляти дані для організацій невеликих або великих масштабів. Це дозволяє поєднувати дані з різних джерел для створення загального сховища та додатково вдосконалює дані для створення розгалужених звітів, які діляться між відділами для звітування. [47]

Apache Mahout - проект, розроблений Фондом Apache, який служить основною метою створення алгоритмів машинного навчання. Основна увага приділяється кластеризації даних, класифікації та спільній фільтрації. Mahout написаний на JAVA і включає бібліотеки JAVA для виконання мате-

матичних операцій, таких як лінійна алгебра та статистика. Mahout постійно зростає, оскільки алгоритми, реалізовані всередині Apache Mahout, постійно зростають. Алгоритми Mahout реалізували рівень над Hadoop за допомогою картографування / зменшення шаблонів.

Mahout має такі основні функції:

Розширене середовище програмування;

Попередньо складені алгоритми;

Середовище експериментів з математикою;

Розрахунки на GPU для підвищення продуктивності. [48]

Deductor Studio є аналітичною платформою – основою для створення закінчених прикладних рішень в області аналізу даних. Реалізовані в Deductor Studio технології дозволяють на базі єдиної архітектури пройти всі етапи побудови аналітичної системи: від створення сховища даних до автоматичного підбору моделей і візуалізації отриманих результатів.

Deductor Studio реалізує функції імпорту, обробки, візуалізації та експорту даних. Вона може функціонувати і без сховища, отримуючи інформацію з будь-яких інших джерел, але найбільш оптимальним є їх спільне використання. В Deductor Studio включено повний набір механізмів, що дозволяють отримати інформацію з довільного джерела даних, провести весь цикл обробки (очищення, трансформацію даних, побудову моделей), відобразити отримані результати найбільш зручним чином (OLAP, таблиці, діаграми, дерева і т. д.) і експортувати їх в найбільш поширені формати. [49].

Таблицю з визначеними перевагами та недоліками приведених програмних продуктів наведено в таблиці 2.1

Таблиця 2.1 – Порівняльні переваги програмних засобів

Назва	Переваги	Недоліки
Statistica	Системні потрібності мінімальні Кількість методі статистичного аналізу Наявність візуального відображення звітів	Відсутність потрібного методу DataMining
Orange	Доступні різні компоненти які реалізовані на мові Python. Велика інтерактивність і більш весела атмосфера. Можливість приймати різні рішення що до подальшого аналізу даних	Потрібні великі розрахункові можливості ПК Відсутність потрібного методу DataMining

Продовження таблиці 2.1

Назва	Переваги	Недоліки
Rapid Miner	Має інтегроване середовище для DataMining Широкий спектр застосування	Відсутність потрібного методу DataMining Потрібність приватного або публічного серверу
Waikato Environment	Виконує основні завдання щодо обробки даних Інтерактивний інтерфейс Програмний продукт написаний на JAVA	Потрібність доступу ПЗ Потрібність даних в вигляді плоского файлу Відсутність потрібного методу DataMining
KNIME	Концепція модульного конвеєра Можливість масштабування Використання вузлів для попередньої обробки даних	Потрібні великі розрахункові можливості ПК Відсутність потрібного методу DataMining
Sisense	Наглядні звіти Можливість обробляти дані різного масштабу Розгалужені звіти	Відсутність потрібного методу DataMining
Apache Mahout	Кластеризація даних Можливість підключати бібліотеки JAVA Розрахунки за допомогою GPU	Відсутність потрібного методу DataMining
Deductor Studio	Доступність всіх етапі для побудови аналітичної системи Реалізація потрібних методів DataMining Автоматичне виконання операцій з даними	Потрібність ліцензованого доступу

Провівши детальний аналіз програмних продуктів, можна дійти висновку, реалізація поставлених у дослідженні задач найкращим чином може бути здійснена за допомогою є Deductor Studio, оскільки саме в цьому програмному засобі є можливість використовувати потрібні для дослідження методи DataMining - пошук асоціативних правил і побудова самоорганізованих карт

РОЗДІЛ 3 СТАТИСТИЧНО-ЙМОВІРНІСНА ОЦІНКА ЗДІЙСНЕННЯ КОРУПЦІЙНИХ ЗЛОЧИНІВ В УКРАЇНІ

3.1 Налаштування вхідних даних

Для дослідження у роботі було використано базу даних про корупційні правопорушення, що були здійснені в Україні за період з 2012 року до третього кварталу 2018 року. У базі, яка представляє собою таблицю в форматі Excel знаходиться понад 26 тис. рядків з записами про кримінальні справи та вироки по корупційним правопорушення. Слід відзначити, що і є інформація по корупційним злочинам у Донецькій та Луганській областях за 2012-2013 роки у повному обсязі і з 2014 по 2018 роки лише по тих частинах областей, що не знаходяться на території ОРДЛО.

База даних містить наступну інформацію (це змінні, значення яких розміщено у стовпцях таблиці) про корупційні злочини:

– дата реєстрації справи (REG_DATE);

– реєстраційний номер справи(REG_NUM);

– вирок (PUNISHMENT);

– ПІБ (FIO);

– місце роботи (JOBPLACE);

– посада (JOBPOST);

– стаття кримінального кодексу (CODEXARTICLE);

– дата винесення вироку (JUDGMENTDATE);

– номер вироку (JUDGMENTNUMBER).

Фрагмент початкових даних у вигляді таблиці наведено в додатку Б.

Така організація даних для проведення досліджень неприпустима і є надлишковою. Більш того, для побудови асоціативних правил, у якості ознаки транзакції буде обрано номер справи. Однак у деяких рядках таблиці з

даними у розділі «Стаття» приведено декілька статей одночасно. Тому для правильної та коректної роботи було вирішено сформувати дані таким чином, щоб у одному рядку за одним номером справи було вказано лише одну статтю вироку. Інші статті за одним вироком заносились в нові додані рядки з дублюванням інформації по справі. Це призвело до того, що у таблицю було додано понад 1000 нових рядків.

Додатково у таблиці з даними було проведено узагальнення:

– по займаним посадам (новий стовпчик «Статус»): керівник, службовець, підприємець, депутат, інші;

– по сфері діяльності (новий стовпчик «Галузь»): держслужбовці, медичні працівники, освіта, ЗСУ, органи самоврядування, підприємці, працівники фіскальних органів, фінансові працівники, юристи та інші.

З даних, що були в нашому розпорядженні, для розрахунків були потрібні наступні:

– номер справи;

– номер статті за вироком;

– займана посада (статус);

– сфера діяльності (галузь).

Екранний фрагмент остаточного варіанту організації вхідних даних представлено на рис. 3.1.

	Столбец3	Статус	Стаття2	Галузь	Столбец2
2	45686_	Керівник	15\	Держслужбовець	КерівникДержслужбовець
3	45686_	Керівник	19\	Держслужбовець	КерівникДержслужбовець
4	45710_	Службовець	15\	Держслужбовець	СлужбовецьДержслужбовець
5	45710_	Службовець	27\	Держслужбовець	СлужбовецьДержслужбовець
6	46004_	Службовець	15\	Освіта	СлужбовецьОсвіта
7	46004_	Службовець	27\	Освіта	СлужбовецьОсвіта
8	47148_	Керівник	15\	Держслужбовець	КерівникДержслужбовець
9	47148_	Керівник	307\	Держслужбовець	КерівникДержслужбовець
10	49158_	Керівник	28\	Приватний підприємиць	КерівникПриватний підприємиць
11	49158_	Керівник	19\	Приватний підприємиць	КерівникПриватний підприємиць
12	45679_	Службовець	191\	Держслужбовець	СлужбовецьДержслужбовець
13	45679_	Службовець	27\	Держслужбовець	СлужбовецьДержслужбовець
14	47150_	Керівник	115\2	Приватний підприємиць	КерівникПриватний підприємиць
15	15311800_	Службовець	12\1	Фіскальний орган	СлужбовецьФіскальний орган
16	49306_	Керівник	122\1	Приватний підприємиць	КерівникПриватний підприємиць
17	15302059_	Службовець	14\	Місцеве самоврядування	СлужбовецьМісцеве самоврядування
18	15302059_	Службовець	14\	Місцеве самоврядування	СлужбовецьМісцеве самоврядування
19	15296169_	Службовець	14\	Місцеве самоврядування	СлужбовецьМісцеве самоврядування
20	15296169_	Службовець	14\	Місцеве самоврядування	СлужбовецьМісцеве самоврядування

Рисунок 3.1. – Фрагмент вхідних даних

Подальші розрахунки проводились в Deductor Studio, який дозволяє в якості джерела даних використовувати у тому числі і файли MS Excel.

3.2 Аналіз кримінальних право порушень за допомогою карт Кохонена

Для побудови карт Кохонена і формування кластерів було обрано три показники: інформація про займану посаду (статус), про сферу діяльності (галузь) та інформація про статтю вироку (Стаття2). Етапи налаштування карт були наступні:

1. Вибір вхідних змінних (рис. 3.2.)

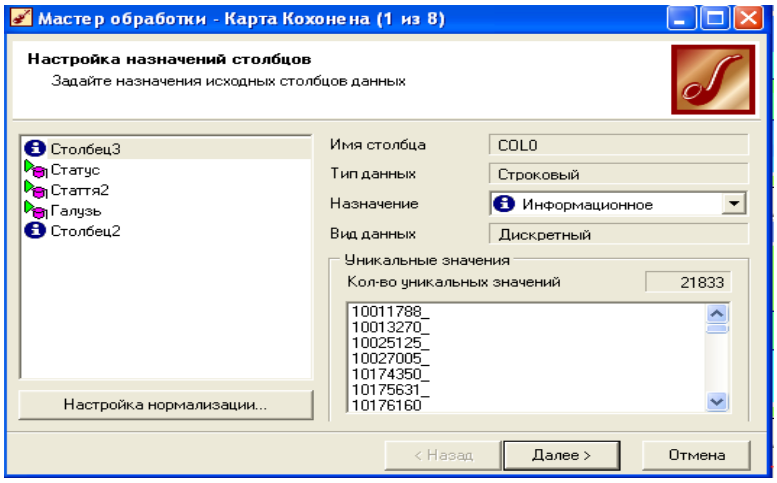


Рисунок 3.2 – Перший етап налаштування побудови карт Кохонена

2. Визначення обсягів навчальної та тестової вибірок (рис. 3.3)

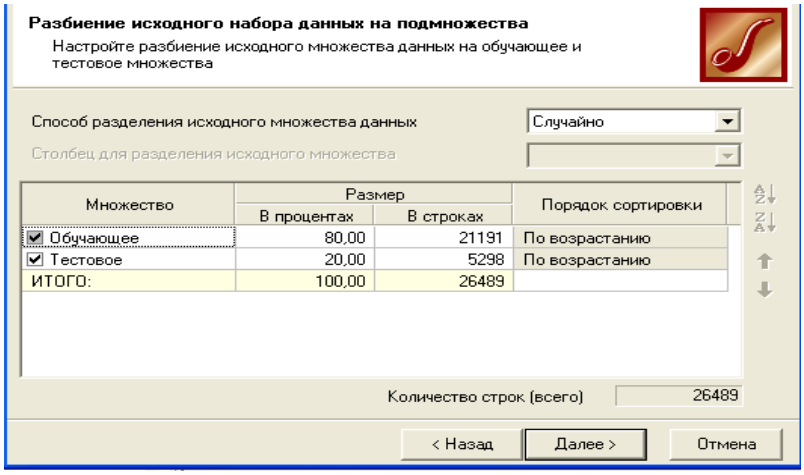


Рисунок 3.3 – Другий етап налаштування побудови карт Кохонена

3. Визначення розмірності площини побудови карт Кохонена та форми комірок (рис. 3.4)

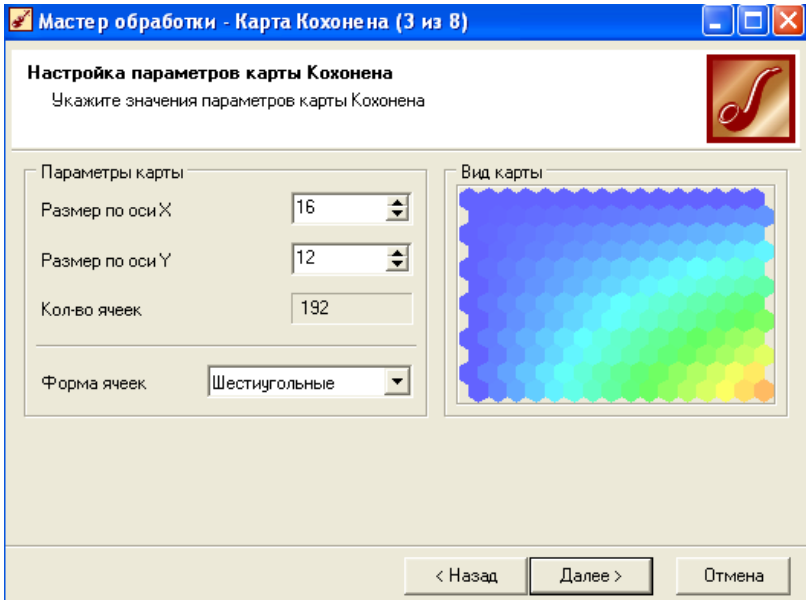
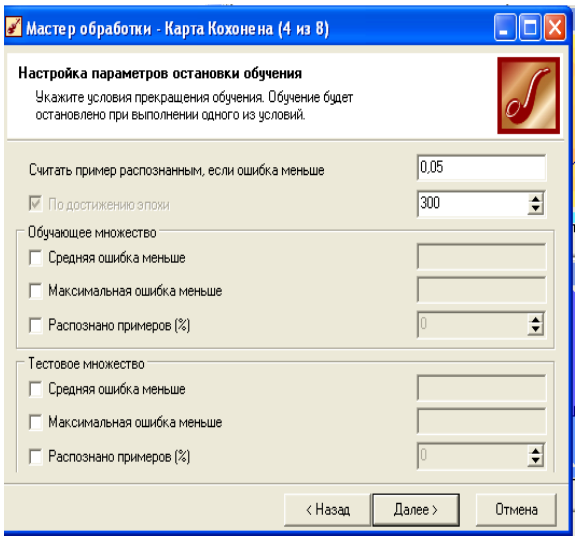
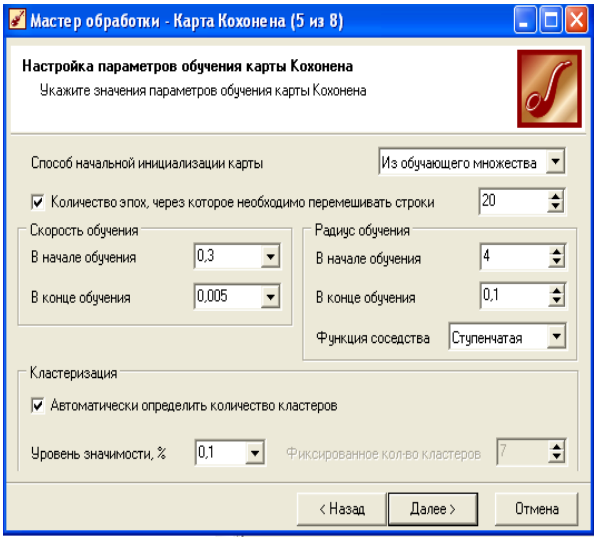


Рисунок 3.4 – Третій етап налаштування побудови карт Кохонена

4. Встановлення параметрів навчання нейромережі (рис. 3.5а, б)

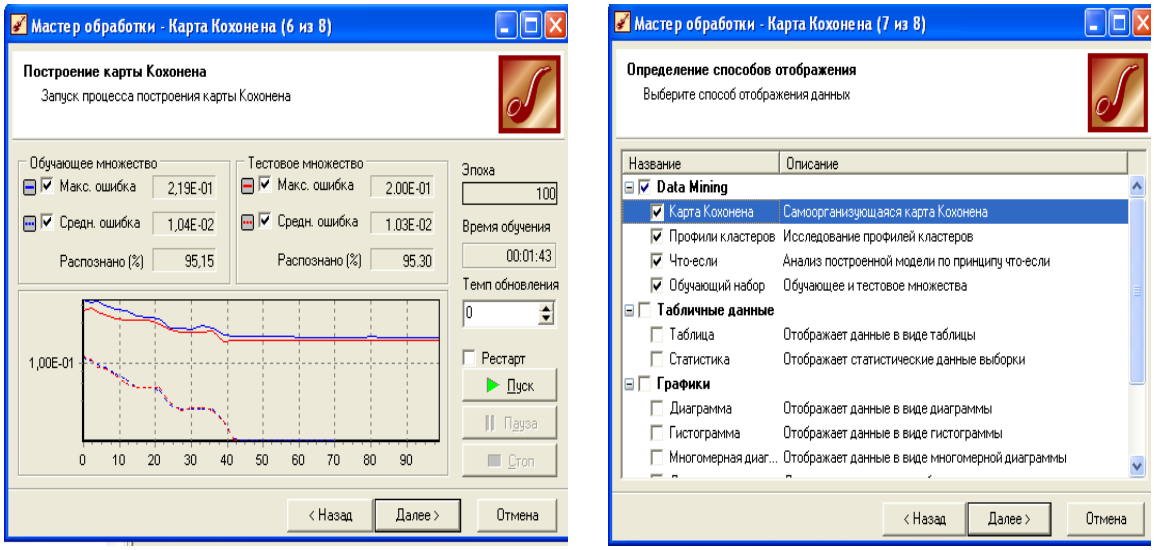




а) б)

Рисунок 3.5 – Четвертий (а) та п’ятий (б) етапи процесу налаштування побудови карт Кохонена

Після налаштування необхідних параметрів відбувається безпосередньо процес навчання (300 епох) та визначення способу відображення отриманих результатів (рис. 3.6 а, б).



а) б)

Рисунок 3.6 – Шостий (а) та сьомий (б) етапи побудови карт Кохонена

В результаті налаштувань дані було розподілено по шести кластерам. На рисунках 3.7 – 3.9 приведено розподіл елементів по кластерам у розрізах «статус», «стаття» та «галузь». Межі кластерів на картах виділено чорним кольором.

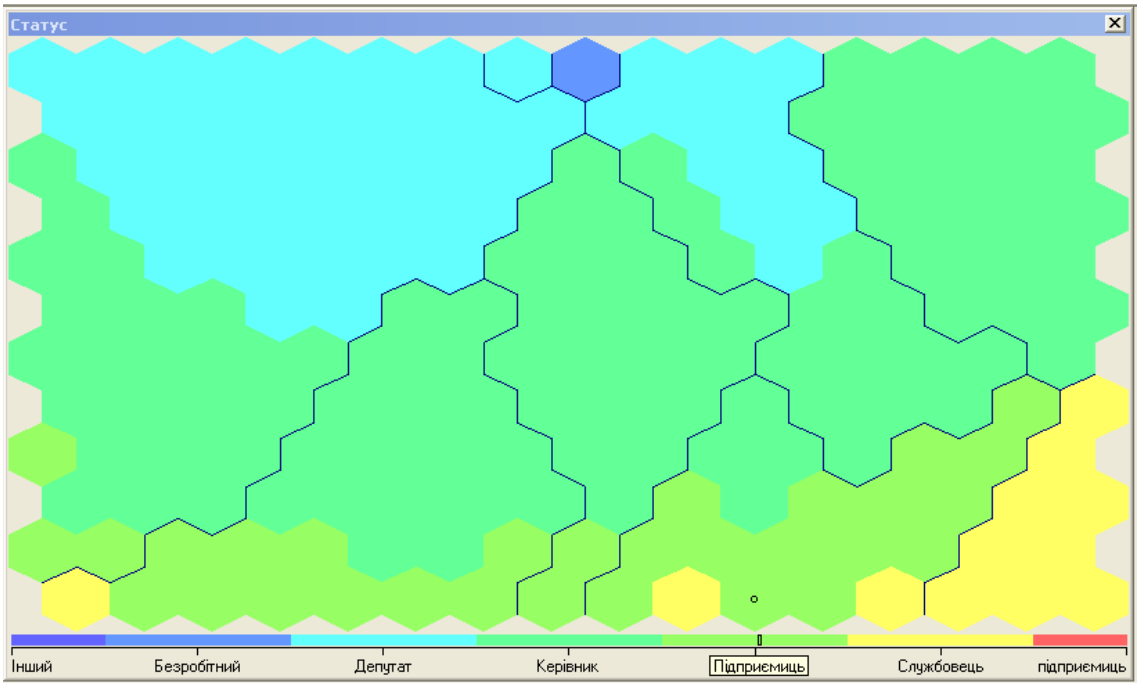


Рисунок 3.7 – Розподіл даних по кластерам у розрізі «статус»

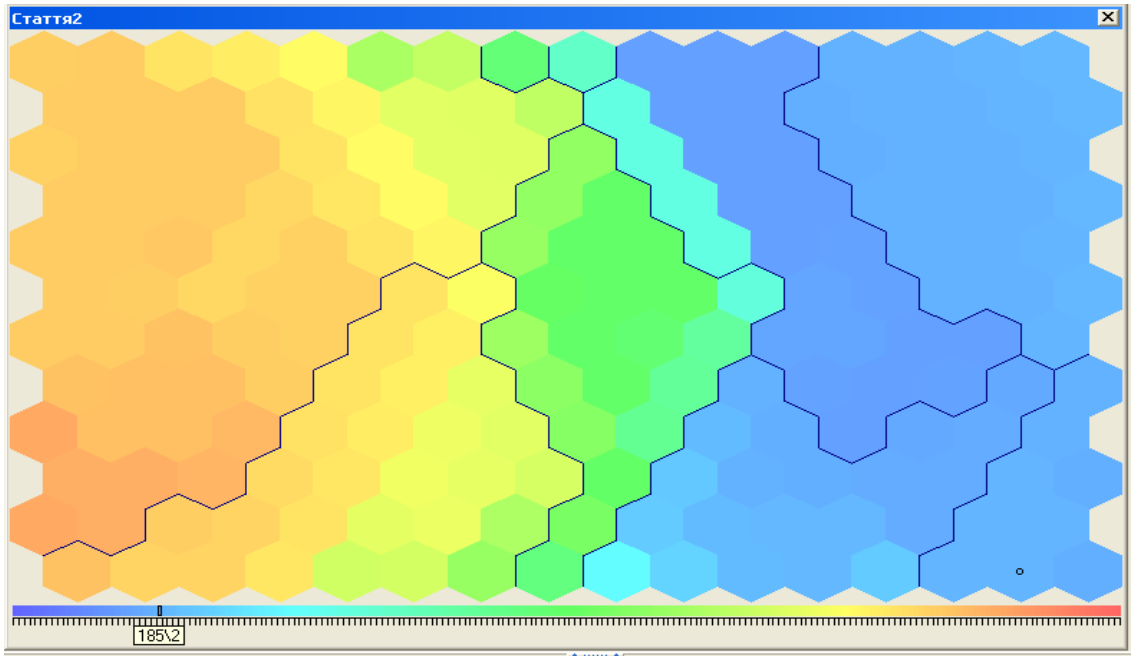


Рисунок 3.8 – Розподіл даних по кластерам у розрізі «стаття»

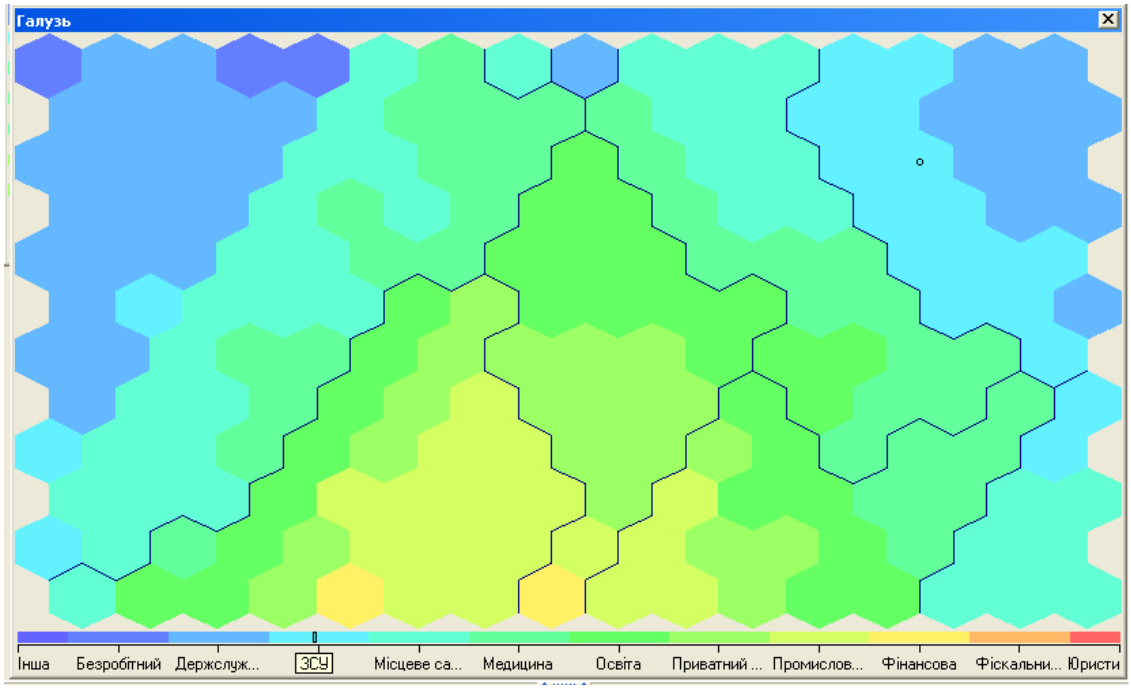


Рисунок 3.9 – Розподіл даних по кластерам у розрізі «галузь»

За результатами аналізу побудованих карт Кохонена (рис. 3.7-3.9) можна зробити деякі узагальнюючі висновки щодо зв’язку кримінальної статті корупційного злочину, сфери діяльності людини та посади, яку займала така людина.

Наприклад, якщо людина працює в освіті на керівній посаді, то існує велика ймовірність, що корупційний злочин може бути здійснено за наступними статтями:

стаття 27 – «Види співучасників»;

стаття 185 – «Крадіжка»;

стаття 366 – «Службове підроблення».

Якщо ж сферою діяльності людини є ЗСУ і вона працює на командній (керівній) посаді, то існує велика ймовірність, що корупційний злочин може бути здійснено за наступними статтями:

стаття 185 – «Крадіжка»;

стаття 369 – «Пропозиція обіцянка, надання неправомірної вигоди службовій особі».

Якщо людина працює в органах місцевого самоврядування на керівній посаді, то існує велика ймовірність, що корупційний злочин може бути здійснено за наступними статтями:

стаття 185 – «Крадіжка»;

стаття 368 – «Незаконне збагачення».

Якщо людина працює в органах місцевого самоврядування і є при цьому депутатом, то існує ймовірність, що корупційний злочин може бути здійснено за наступними статтями:

стаття 172 – «Грубе порушення законодавства про працю»;

стаття 364 – «Зловживання владою або службовим становищем»;

стаття 362 – «Несанкціоновані дії з інформацією».

Якщо особа працює на керівній посаді в приватному підприємстві, існує ймовірність, що корупційний злочин може бути здійснено за наступними статтями:

стаття 27 – «Види співучасників»;

стаття 240 – «Порушення правил охорони та використання надр»;

стаття 307 – «Незаконне виробництво;

стаття 364 – «Зловживання владою або службовим становищем».

Подібні результати було отримано у розрізах займаних посад та сфер діяльності. Для збереження місця ми їх не приводимо, а отримати цю інформацію можна в Deductor Studio, оскільки побудовані карти є інтерактивними.

3.3 Асоціативні правила по виявленню корупційних ризиків в Україні

Для визначення асоціативних правил у роботі виходили з наступного припущення: вироки за корупційні злочини деяким особам одночасно оголошувались одночасно за кількома кримінальними статтями у рамках однієї справи. Тому ознакою транзакції було обрано номер кримінальної справи (ID). Набором значень показника (елементу) за транзакцією виступають номери статті корупційного правопорушення – (Стаття2). Діалогове вікно вибору транзакції та елементів транзакції приведено на рис. 3.10.

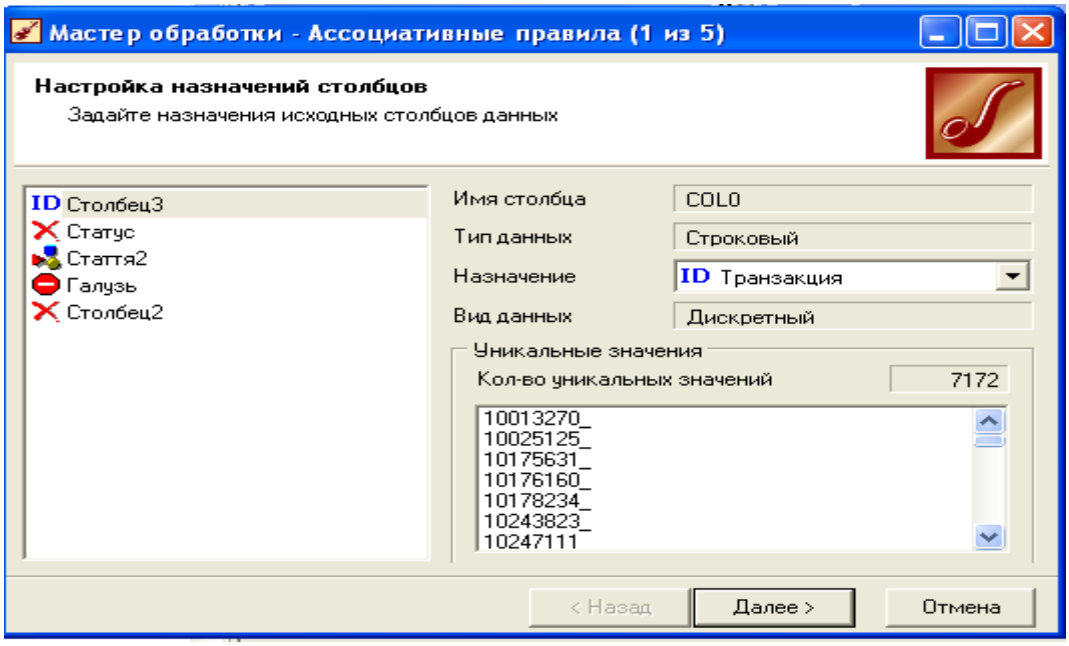


Рисунок 3.10 – Перший етап налаштування пошуку асоціативних правил

На наступному кроці необхідно виконати налаштування кожного з чотирьох значень параметрів для пошуку правил:

Мінімальна і максимальна підтримка правила. Асоціативні правила знаходяться тільки в деякій множині всіх транзакцій. Для того щоб певна транзакція увійшла в таку множину, вона повинна зустрічатись у вхідній вибірці кількість разів, більше мінімальної підтримки і менше максимальної.

Мінімальна і максимальна достовірність. Це процентне відношення кількості транзакцій, що містять всі елементи, які входять в правило, до кількості транзакцій, що містять елементи, які входять в умову. Якщо транзакція – це номер справи, а елемент – стаття, то достовірність характеризує, наскільки часто зустрічається така стаття вироку, якщо вирок вже було оголошено за іншими статтями цього ж вироку для всіх вироків, що потрапили у таке правило. [50] Вікно налаштувань параметрів асоціативних правил, мінімальні та максимальні значення приведено на рис. 3.11.

Подальші етами побудови дозволяють визначити кількість правил та кількість множин з асоціативними правилами, а також спосіб відображення результатів (рис. 3.12а, б). На нашу думку, найбільш цікаві результати можуть бути подані у вигляді безпосередньо списку правил (умова – «стаття», наслідок «супутня стаття») та дерев правил (умова – «стаття», наслідок «супутня стаття» з групуванням по статтям).

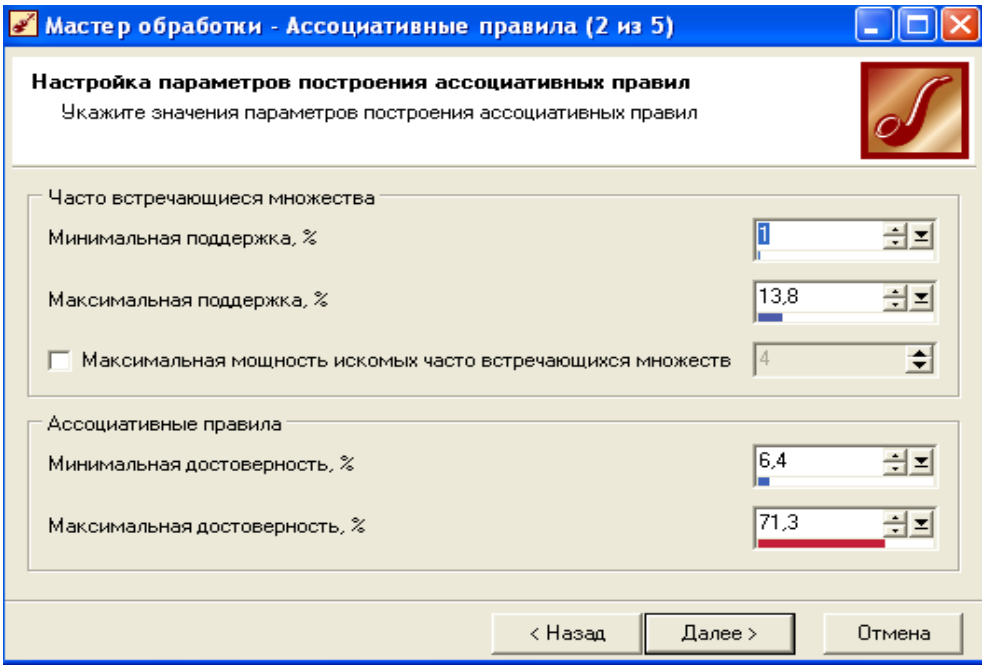
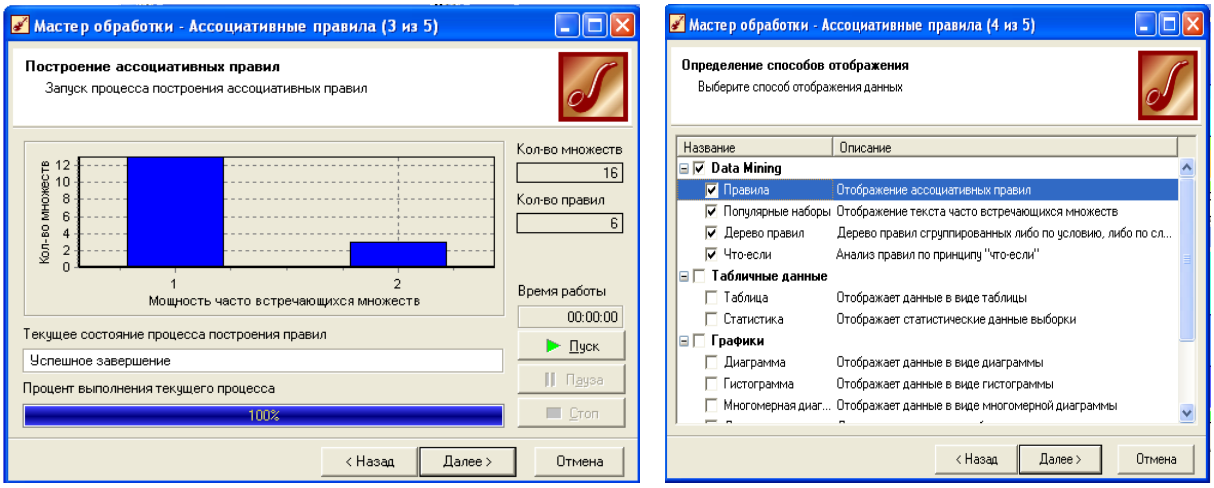


Рисунок 3.11 – Другий етап налаштування пошуку асоціативних правил



а) б)

Рисунок 3.12 – Третій (а) та четвертий (б) етапи налаштування пошуку асоціативних правил

Для більш точного визначення ймовірності здійснення корупційних злочинів за допомогою асоціативних правил було вирішено шукати такі правила окремо за статусом (посадою) та сферою діяльності (галузь) осіб, що здійснили правопорушення. Для реалізації поставленої задачі перед налаштуванням правил (до початку першого етапу) здійснювалось фільтрування вхідних даних за кожним можливим значенням у категоріях «галузь» та «посада». Таким чином було отримано 15 груп правил. Всі ці групи приведено в Додатку В та Додатку Г роботи.

Для прикладу проаналізуємо асоціативна правила, побудовані для держслужбовців, працівників органів місцевого самоврядування.

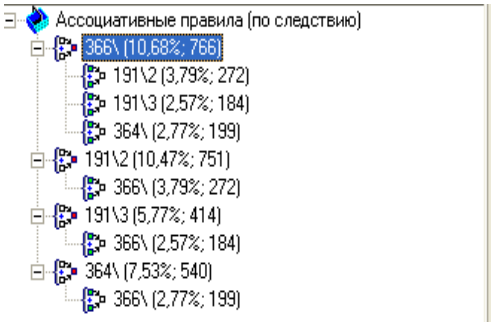
Так, для категорії держслужбовців (рис. 3.13), було знайдено шість найбільш ймовірних правил щодо можливості здійснення корупційних злочинів:

Якщо злочин було здійснено за статтею 366 - «Службове підроблення», з ймовірністю 36% можна очікувати, що буде здійснено правопорушення за статтею 191 ч.2 – «Привласнення, розтрата майна або заволодіння ним шляхом зловживання службовим становищем» або з ймовірністю 44% за статтею 191 ч. 3 – «Привласнення, розтрата майна або заволодіння ним шля-

хом зловживання службовим становищем» або з ймовірністю майже 37% за статтею 364 – «Зловживання владою або службовим становищем».

Правил: 6 из 6 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	366\	272	3,79	36,22	3,391
2	2	366\	191\2	272	3,79	35,51	3,391
3	3	191\3	366\	184	2,57	44,44	4,161
4	4	366\	191\3	184	2,57	24,02	4,161
5	5	364\	366\	199	2,77	36,85	3,450
6	6	366\	364\	199	2,77	25,98	3,450



Количество правил: 3; Следствие: 366\

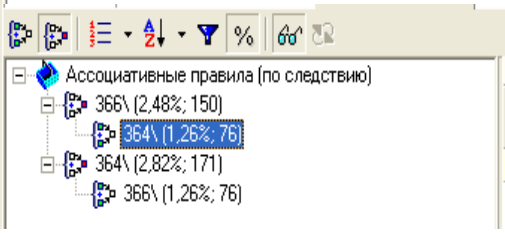
Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	272	3,79	36,20	3,391
191\3	184	2,57	44,40	4,161
364\	199	2,77	36,90	3,45

Рисунок 3.13 – Асоціативні правила та дерева правил для категорії «держслужбовці»

Для категорії працівників органів місцевого самоврядування було знайдено два правила щодо ймовірності здійснення корупційних злочинів (рис. 3.14).

Правил: 2 из 2 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	364\	366\	76	1,26	44,44	17,938
2	2	366\	364\	76	1,26	50,67	17,938



Правило №1; Следствие: 366\

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
364\	76	1,26	44,40	17,938

Рисунок 3.14 – Асоціативні правила та дерева правил для категорії «працівники органів місцевого самоврядування»

З представлених результатів випливає, працівниками даної категорії найбільш ймовірними є корупційні злочини, що здійснюються за статтями

366 – «Службове підроблення» статті і 364 – «Зловживання владою або службовим становищем», причому якщо кримінальну справу було відкрито за статтею 364, існує ймовірність у 44%, що за цією ж справою буде знайдено правопорушення за статтею 366. В той де час, якщо кримінальну справу було відкрито за статтею 366, то майже на 50% існує ймовірність, що у межах кримінальної справи буде відкрите провадження за статтею 364.

Дещо інші результати можна отримати аналізуючи сфери діяльності, у яких було здійснено корупційні злочини. Наприклад, для категорії «Медицина» (рис. 3.15) було отримано наступні результати.

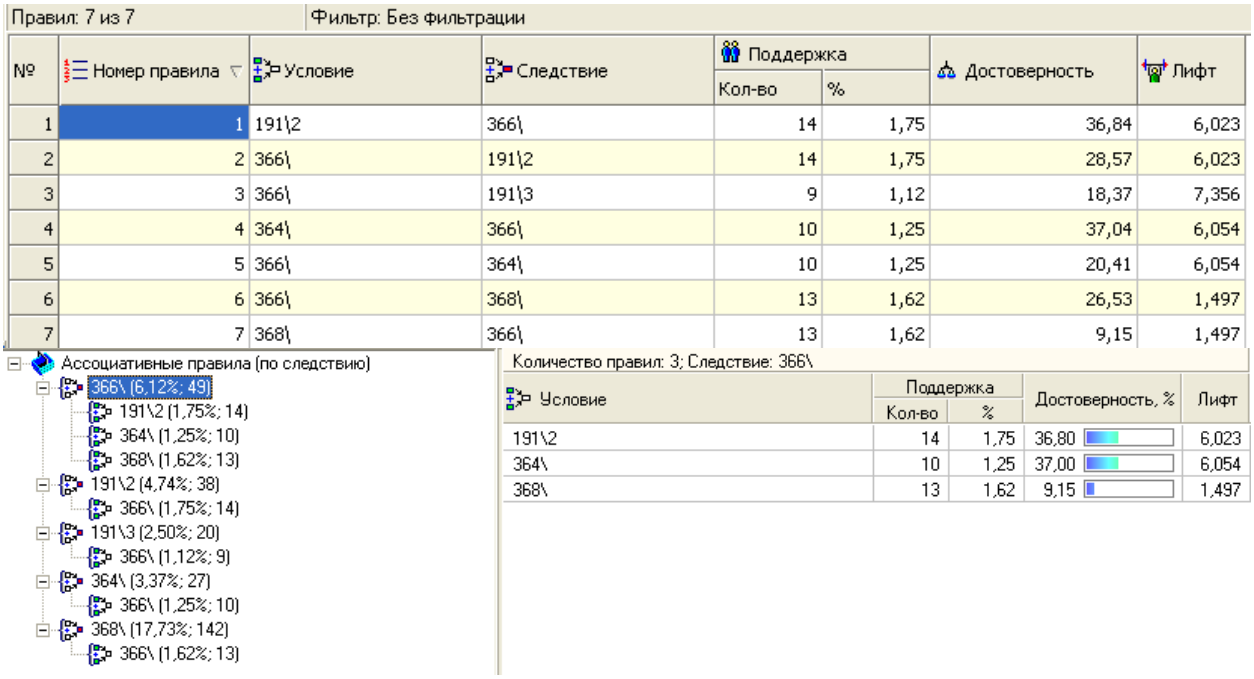


Рисунок 3.15 – Асоціативні правила та дерева правил для категорії «медицина»

Всього було знайдено 7 правил. Якщо злочин було здійснено за статтею 366 - «Службове підроблення», то з ймовірністю 37% у межах справи міг бути злочин і за статтею 191 ч.2 - «Привласнення, розтрата майна або заволодіння ним шляхом зловживання службовим становищем» або за статтею 364 - «Зловживання владою або службовим становищем» або ж з ймовірністю 9% за статтею 368 – «Прийняття пропозиції, обіцянки або одержання службовою особою неправомірної вигоди».

ВИСНОВКИ

Таким чином при написанні дипломної роботи було вирішено наступні задачі.

Досліджено поняття корупції і та основні підходи щодо визначення корупційних злочинів.

Проведено статистичний аналіз кримінальних справ з корупційних злочинів в Україні за 2012-2018 роки та визначено методи статистично-ймовірнісної оцінки виявлення корупційної діяльності.

У роботі було визначено можливість застосування методів DataMinig для розрахунку статистично-ймовірнісної оцінки корупційних ризиків завдяки орієнтації цих методів на роботу в великими обсягами даних, а також наявності відповідного програмного забезпечення, що у свою чергу дозволяє автоматизувати процеси обробки даних.

У роботі було проведено порівняльний аналіз програмного забезпечення для автоматизації розрахунків та обгрунтовано вибір найбільш відповідного програмного продукту – Deductor Studio.

Для реалізації поставленої у роботі мети було побудовано карти Кохонена та асоціативні правила по корупційним справам в Україні та проаналізовано отримані результати.

Результати аналізу дозволяють оцінювати ймовірність скоєння корупційних злочинів залежно від сфери діяльності та посади можливих злочинців, а також ймовірність того, що при скоєнні певного корупційного злочину додатково може бути скоєно інший корупційний злочин.

На нашу думку, отримані результати можуть бути використані у практичній діяльності як контролюючих, так і правоохоронних органів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Corruption Perceptions Index 2018. URL: https://www.transparency.org/cpi2018/results – Назва з екрана (дата звернення 09.12.2019)

2. Закон Про боротьбу з корупцією Про запобігання корупції. URL: https://zakon.rada.gov.ua/laws/show/356/95-%D0%B2%D1%80/ed20110101/find?text=%EA%EE%F0%F3%EF%F6%B3%BA%FE – Назва з екрану (дата звернення 09.12.2019)

3. Morris, S.D. (1991), Corruption and Politics in Contemporary Mexico. University of Alabama Press, Tuscaloosa P. 224.

4. Senior, I. (2006), Corruption – The World's Big C., Institute of Economic Affairs, London P. 112.. URL: https://iea.org.uk/wp-content/uploads/2016/07/upldbook324pdf.pdf (дата звернення 09.12.2019)

5. Kaufmann, Daniel; Vicente, Pedro (2005). «Legal Corruption» . World Bank. Archived from the original (PDF) on 5 May 2015. Retrieved 25 September 2012. P. 43.. URL: http://siteresources.worldbank.org/INTWBIGOVANTCOR/Resources/Legal_Corruption.pdf (дата звернення 09.12.2019)

6. Mishler v. State Bd. of Med. Examiners». Justia Law.. URL: https://law.justia.com/cases/nevada/supreme-court/1993/22397-1.html – Назва з екрану. (дата звернення 09.12.2019)

7. Supreme Court Of The United States Report». P. 36.. URL: https://www.supremecourt.gov/opinions/14pdf/13-534_19m2.pdf – Назва з екрану. (дата звернення 09.12.2019)

8. «Material on Grand corruption» (PDF). United Nations Office on Drugs and Crime.. URL: https://www.unodc.org/documents/NGO/Grand_Corruption_definition_with_explanation_19_August_2016_002_1.pdf – Назва з екрану. (дата звернення 09.12.2019)

9. Alt, James. «Political And Judicial Checks On Corruption: Evidence From American State Governments». Projects at Harvard.. URL: https://web.archive.org/web/20151203043655/http:/projects.iq.harvard.edu/gov2126/files/altlassen.pdf (дата звернення 09.12.2019)

10. «Glossary». U4 Anti-Corruption Resource Centre. Retrieved 26 June 2011.. URL: https://www.u4.no/terms# (дата звернення 09.12.2019)

11. Lorena Alcazar, Raul Andrade (2001). Diagnosis corruption. P. 135–136.

12. Znoj, Heinzpeter (2009). «Deep Corruption in Indonesia: Discourses, Practices, Histories». In Monique Nuijten, Gerhard Anders Corruption and the secret of law: a legal anthropological perspective. Ashgate. P. 53–54.

13. Legvold, Robert (2009). «Corruption, the Criminalized State, and Post-Soviet Transitions». In Robert I. Rotberg Corruption, global security, and world orde. Brookings Institution. P. 197.

14. Merle, Jean-Christophe, ed. (2013). «Global Challenges to Liberal Democracy». Spheres of Global Justice. 1:P. 812.

15. Pogge, Thomas. «Severe Poverty as a Violation of Negative Duties». URL: https://web.archive.org/web /20150211060650/http:/thomaspogge.com/publications-on-global-justice/ (дата звернення 09.12.2019)

16. R. Keith Schoppa, Revolution and Its Past: Identities and Change in Modern Chinese History P. 383.

17. Хабарництово - тлумачення орфографія правопис. URL: https://www.slovnyk.ua/index.php?swrd=%D0%A5%D0%B0%D0%B1%D0%B0%D1%80%D0%BD%D0%B8%D1%86%D1%82%D0%B2%D0%BE – Назва з екрану. (дата звернення 09.12.2019)

18. Відкат - тлумачення орфографія правопис. URL: https://www.slovnyk.ua/index.php?swrd=%D0%92%D0%86%D0%94%D0%9A%D0%90%D0%A2 – Назва з екрану. (дата звернення 09.12.2019)

19. Шахрайство - тлумачення орфографія правопис. URL: https://www.slovnyk.ua/index.php? swrd=if%5Bhfqcndj – Назва з екрану. (дата звернення 09.12.2019)

20. Про запобігання корупції. URL: https://zakon.rada.gov.ua/laws/main/1700-18 – Назва з екрану. (дата звернення 09.12.2019)

21. Змова - тлумачення орфографія правопис. URL: https://www.slovnyk.ua/index.php?swrd=%D0%B7%D0%BC%D0%BE%D0%B2%D0%B0 – Назва з екрану. (дата звернення 09.12.2019)

22. Картель - тлумачення орфографія правопис. URL: https://www.slovnyk.ua/index.php?swrd=%D0%BA%D0%B0%D1%80%D1%82%D0%B5%D0%BB%D1%8C – Назва з екрану. (дата звернення 09.12.2019)

23. Intersection of business and politics: Problem of the ‘revolving door’ in Georgia. URL: https://transparency.ge/en/blog/intersection-business-and-politics-problem-revolving-door-georgia – Назва з екрану. (дата звернення 09.12.2019)

24. Фаворитизм - тлумачення орфографія правопис. URL: https://www.slovnyk.ua/index.php?swrd=%D0%A4%D0%B0%D0%B2%D0%BE%D1%80%D0%B8%D1%82%D0%B8%D0%B7%D0%BC – Назва з екрану. (дата звернення 09.12.2019)

25. Інсайдерська торгівля - тлумачення орфографія правопис. URL: https://dic.academic.ru/dic.nsf/fin_enc/32813 – Назва з екрану. (дата звернення 09.12.2019)

26. Кримінальний Кодекс України .. URL: https://zakon.rada.gov.ua/laws/show/2341-14 – Назва з екрану. (дата звернення 09.12.2019)

27. Качиньський А. Б. Безпека, загрози і ризик: наукові концепції та матетаматичні методи / А.Б. качиньський. – К.: Поліграфконсалдинг, 2004. – 472 с.

28. Guide for Anti-Corruption Risk Assessment. URL: http://www.cop-advanced.org/sites/default/files/docs/RESSOURCES/Lutte_contre_la_corruption/AGuideforAntiCorruptionRiskAssessment.pdf – Назва з екрану (дата звернення 09.12.2019)

29. «DataMining Curriculum». ACM SIGKDD. . URL: https://www.kdd.org/curriculum/index.html – Назва з екрану.

30. Clifton, Christopher «Encyclopædia Britannica: Definition of DataMining». .. URL: https://www.britannica.com/technology/data-mining – Назва з екрану

31. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). «From DataMining to Knowledge Discovery in Databases». .. URL: https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf – Назва з екрану (дата звернення 09.12.2019)

32. Han, Jiawei; Kamber, Micheline (2001). DataMining: concepts and techniques. Morgan Kaufmann. P. 740. URL: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf – Назва з екрану (дата звернення 09.12.2019)

33. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). «WEKA Experiences with a Java open-source project». Journal of Machine Learning Research.11:P. 253–254.

34. Olson, D. L. (2007). DataMining in business services. Service Business, 1(3), P. 181-193.

35. Agrawal, R.; Imieliński, T.; Swami, A. (1993). «Mining association rules between sets of items in large databases». Proceedings of the 1993 ACM SIGMOD international conference on Management of data - P. 207.

36. Apriori - масштабований алгоритм пошуку асоціативних правил. URL: https://basegroup.ru/community /articles/apriori – Назва з екрану. (дата звернення 09.12.2019)

37. Complete guide to Association Rules. URL: https://towardsdatascience.com/complete-guide-to-association-rules-2-2-c92072b56c84 – Назва з екрану. (дата звернення 09.12.2019)

38. Kohonen, Teuvo (1982). «Self-Organized Formation of Topologically Correct Feature Maps». Biological Cybernetics - P. 59–69.

39. Abhinav, Ralhan (2018) Self Organizing Maps. URL: https://towardsdatascience.com/self-organizing-maps-ff5853a118d4 – Назва з екрану.

40. Самоорганізовані карти Кохонена - математичний апарат. URL: https://basegroup.ru/community/articles/som – Назва з екрану. (дата звернення 09.12.2019)

41. T. Kohonen, Self-Organizing Maps (Third Extended Edition), New York, 2001, P. 501.

42. STATISTICA Base. URL: http://statsoft.ru/products/statistica_Base/ – Назва з екрану. (дата звернення 09.12.2019)

43. Orange DataMining. URL: https://orange.biolab.si/#Orange-Features – Назва з екрану. (дата звернення 09.12.2019)

44. Rapid Miner. URL: https://rapidminer.com/us/ – Назва з екрану. (дата звернення 09.12.2019)

45. Waikato Environment. URL: https://www.waikatoregion.govt.nz/ – Назва з екрану. (дата звернення 09.12.2019)

46. KNIME. URL: https://www.knime.com/about – Назва з екрану. (дата звернення 09.12.2019)

47. Sisense. URL: https://www.sisense.com/product/ – Назва з екрану. (дата звернення 09.12.2019)

48. Apache Mahout. URL: https://mahout.apache.org/ – Назва з екрану. (дата звернення 09.12.2019)

49. Deductor Studio. URL: https://basegroup.ru/deductor/description – Назва з екрану. (дата звернення 09.12.2019)

50. Асоціативні правила. URL: https://basegroup.ru/deductor/function/algorithm/association-rules – Назва з екрану – Назва з екрану. (дата звернення 09.12.2019)

ДОДАТКИ

Додаток А

SUMMARY

Drozd S. A. statistically probabilistic assessment of corruption risks in Ukraine-qualification master's work. Sumy state University, Sumy, 2019

In the work the probability of Commission of corruption crimes depending on the sphere of activity and a position of possible criminals is investigated also probability of that at Commission of a certain corruption crime in addition other corruption crime can be committed. The main purpose of this study is to develop a model of statistical and probabilistic assessment of corruption risks in Ukraine.

Keywords: associative rules, corruption, Kohonen cards, offenses.

АНОТАЦІЯ

Дрозд С.А. Статистично-ймовірнісна оцінка корупційних ризиків в Україні – Кваліфікаційна магістерська робота. Сумський державний університет, Суми, 2019 р.

У роботі досліджено ймовірність скоєння корупційних злочинів залежно від сфери діяльності та посади можливих злочинців також ймовірність того, що при скоєнні певного корупційного злочину додатково може бути скоєно інший корупційний злочин. Основною метою цього дослідження є розробка моделі статистично-ймовірнісної оцінки корупційних ризиків в Україні.

Ключові слова: асоціативні правила, корупція, карти Кохонена, правопорушення.

Додаток Б

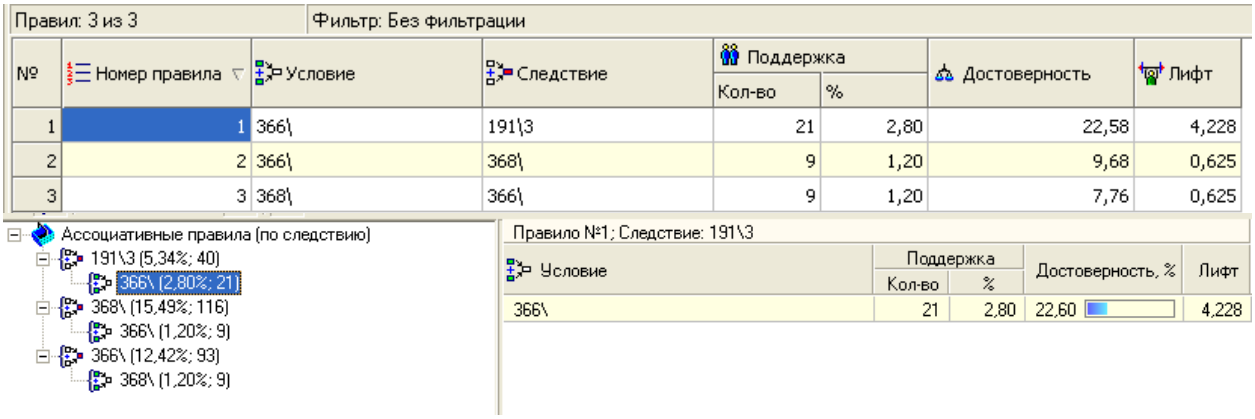
Таблиця з початковими даними

REG_NUM	REG_DATE	PUNISHMENT	FIO	JOBPLACE	JOBPOS	CODEX	CODEXARTICLE	JUDGMENTDATE	JUDGMENTNUMBER
46605	2012-08-17	На підставі ст.70 К	Бондарен	магазин №0302	завідуюч		ККУ, Стаття 366. Службове підроб	2012-03-20	
							ККУ, Частина друга статті 191 Привласнення, розтрата майна або заволодіння ним шляхом зловживання службовим ст		
46606	2012-07-20	Кушнір Людмилу	Кушнір Л	Відділ держав	державний виконав		КУпАП, Стаття 172-7. Порушення	2012-07-11	1514/1021/12
46607	2012-07-25	Згідно ст. 70 ч. 1 К	Гайдейчу	ТзОВ «Експрет	директор		ККУ, Стаття 364. Зловживання вла	2012-07-16	908/1261/2012
							ККУ, Стаття 366. Службове підроблення		
46608	2012-07-26	Османову Аліє Су	Османова	Алупкінська м	спеціалістом 2-ї кат		КУпАП, Стаття 172-6. Порушення	2012-06-12	
46535	2012-03-19	За ч. 2 ст. 190 КК У	Савченко	-	-		КК України, 190 ч. 2 КК України	2011-10-31	1-177/11
							ККУ, Стаття 366. Службове підроблення		
							ККУ, Стаття 365. Перевищення влади або службових повноважень		
							ККУ, Стаття 364. Зловживання владою або службовим становищем		
46538	2012-03-21	За ч. 1 ст. 364 КК У	Денисюк	Козятинська ди	-		ККУ, Стаття 364. Зловживання вла	2012-02-06	1-2/12
							ККУ, Стаття 366. Службове підроблення		
46539	2012-05-04	Визнати винною	Бойчук На	Виборча коміс	Заступник голови ви		КК України, 191 ч.3 КК України	2011-10-17	1-167/11
46540	2012-04-12	На підставі ст. 70 К	Терпелюк	-	-		КК України, 190 ч.2, ст. 27 ч.4 КК	2012-03-14	1-100/12
							ККУ, Стаття 369. Давання хабара		
46545	2012-05-03	Відповідно до ст.(	Рощенко	ТОВ «Нібулон»	агроном		КК України, 191 ч. 3 КК України	2011-12-26	1-197/11
5,00001E+12	2015-12-07	На підставі п. є, ч.	Сугак Вал	МПП «Валеріо	директор		ККУ, Стаття 366. Службове підроб	2014-07-17	524/5763/14-к р.
							ККУ, Частина друга статті 191 Привласнення, розтрата майна або заволодіння ним шляхом зловживання службовим ст		
46546	2012-05-03	На підставі ст.70 ч	Сич Андрі	Прилуцький м	оперуповноважени		КК України, 190 ч.2 та ст.ст. 27 ч.4	2011-07-29	1-258/2011
							ККУ, Стаття 369. Давання хабара		
46547	2012-04-26	На підставі ч. 1 ст.	Бамбуза Г	-	-		КК України, 191 ч.3, ст. 27 КК Укра	2011-12-02	1-206/11
							ККУ, Стаття 358. Підроблення документів, печаток, штампів та бланків, збут чи використання підроблених документі		
46548	2012-04-19	Визнати винною Ф	Факєєва І	Новоселівськи	Начальник відділен		КК України, 191 Привласнення, р	2012-03-01	
46549	2012-04-23	На підставі ст. 75 К	Штанько І	Червонопарти	землевпорядник		ККУ, Стаття 364. Зловживання вла	2011-12-27	1-109/11
46550	2012-03-27	За ст.27 ч.4, ст.191	Малиш М	Путильська ЗО	сезонний кочегар		КК України, 191 ч.3, ст.27 ч.4 КК У	2012-01-30	1-2\12
46551	2012-05-08	У силу ст. 75 КК Ук	Михайлов	Селянське фер	Голова		КК України, 197-1 ч.1 КК України	2012-02-06	1-10/1529/12
							ККУ, Стаття 364. Зловживання владою або службовим становищем		
46552	2012-05-26	Римську Марію Яр	Римська М	Перша міська к	сімейноий лікаря д		КУпАП, Стаття 172-2. Порушення	2011-11-18	3-2552/11
46554	2012-05-30	Адміністративне с	Федорен	Новокрасненс	завідуючий.		КУпАП, Стаття 172-2. Порушення	2012-02-08	
46555	2012-06-01	На підставі ст.75 К	Бабій Пав	Кальненська сі	сільський голова		ККУ, Стаття 364. Зловживання вла	2012-02-03	1908/1-12/12
46556	2012-06-06	Величко Світлану	Величко С	Київський уніе	декан факультету ту		ККУ, Стаття 368-3. Комерційний п	2011-12-14	1-1155/2011
46557	2012-06-06	Адміністративне с	Ващенко	Київський міс	приватний нотаріус		КУпАП, Стаття 172-2. Порушення	2011-10-25	
46558	2012-06-06	На підставі ч. 1 ст.	Бондарчу	Відділення по	начальник.		КК України, 191 3 Привласнення,	2012-03-15	
							ККУ, Стаття 366. Службове підроблення		
46559	2012-06-06	У відповідності до	Гвоздик З	Свірзське лісн	майстер лісу		ККУ, Частина друга статті 191 При	2012-03-16	1-49/2012
46560	2012-06-07	На підставі ст.70 К	Дудік Тет	Виконавчий ко	головний спеціаліст		КК України, 191 3 КК України	2012-04-10	0306/1544/2012
							ККУ, Стаття 365. Перевищення влади або службових повноважень		

Додаток В

Асоціативні правила та дерева правил по галузям

Освіта



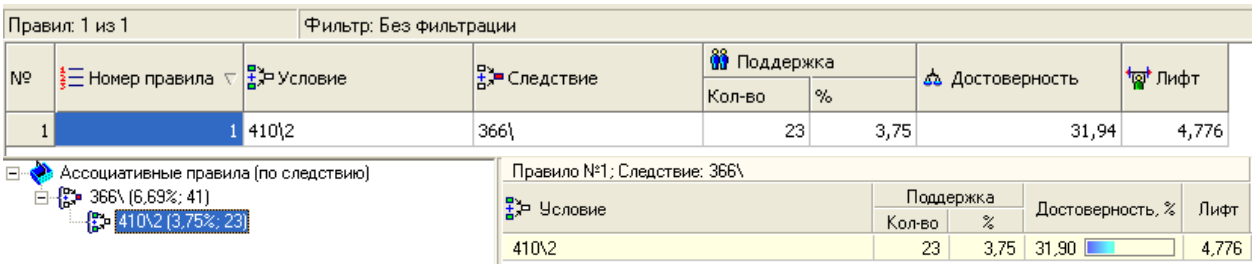
Правил: 3 из 3 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка Кол-во	Поддержка %	Достоверность	Лифт
1	1	366\	191\3	21	2,80	22,58	4,228
2	2	366\	368\	9	1,20	9,68	0,625
3	3	368\	366\	9	1,20	7,76	0,625

Правило №1; Следствие: 191\3

Условие	Поддержка Кол-во	Поддержка %	Достоверность, %	Лифт
366\	21	2,80	22,60	4,228

ЗСУ



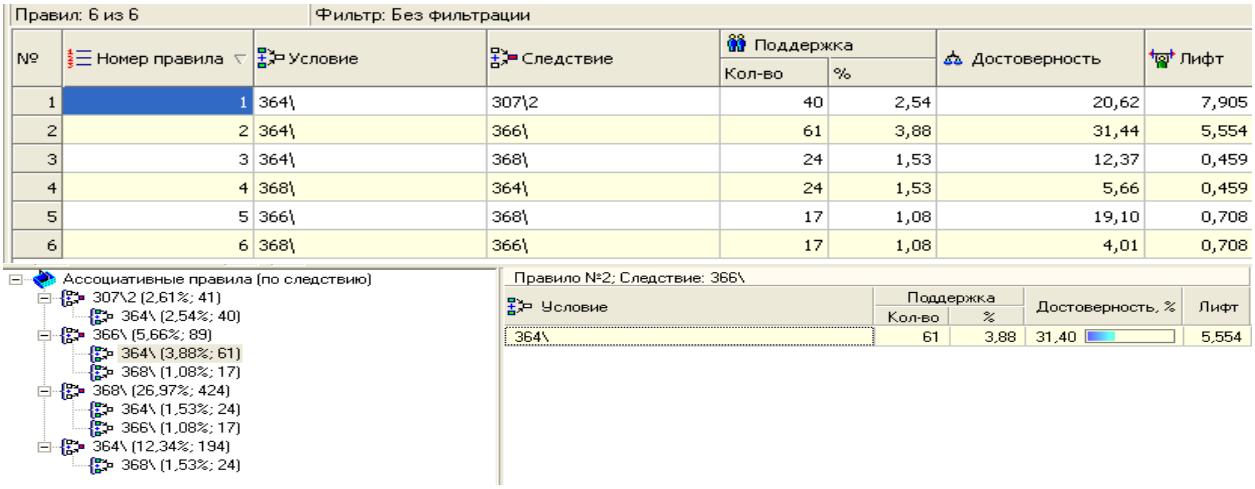
Правил: 1 из 1 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка Кол-во	Поддержка %	Достоверность	Лифт
1	1	410\2	366\	23	3,75	31,94	4,776

Правило №1; Следствие: 366\

Условие	Поддержка Кол-во	Поддержка %	Достоверность, %	Лифт
410\2	23	3,75	31,90	4,776

Фіскальний орган



Правил: 6 из 6 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка Кол-во	Поддержка %	Достоверность	Лифт
1	1	364\	307\2	40	2,54	20,62	7,905
2	2	364\	366\	61	3,88	31,44	5,554
3	3	364\	368\	24	1,53	12,37	0,459
4	4	368\	364\	24	1,53	5,66	0,459
5	5	366\	368\	17	1,08	19,10	0,708
6	6	368\	366\	17	1,08	4,01	0,708

Правило №2; Следствие: 366\

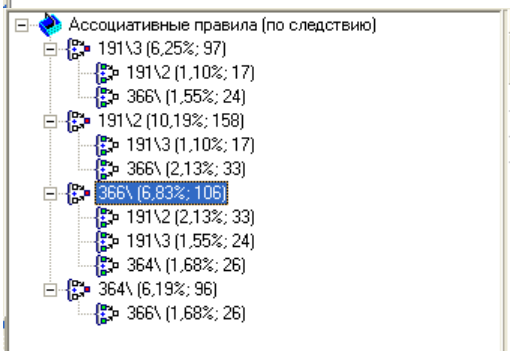
Условие	Поддержка Кол-во	Поддержка %	Достоверность, %	Лифт
364\	61	3,88	31,40	5,554

Продовження додатку В

Промисловість

Правил: 8 из 8 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	191\3	17	1,10	10,76	1,720
2	2	191\3	191\2	17	1,10	17,53	1,720
3	3	191\2	366\	33	2,13	20,89	3,056
4	4	366\	191\2	33	2,13	31,13	3,056
5	5	191\3	366\	24	1,55	24,74	3,620
6	6	366\	191\3	24	1,55	22,64	3,620
7	7	364\	366\	26	1,68	27,08	3,963
8	8	366\	364\	26	1,68	24,53	3,963



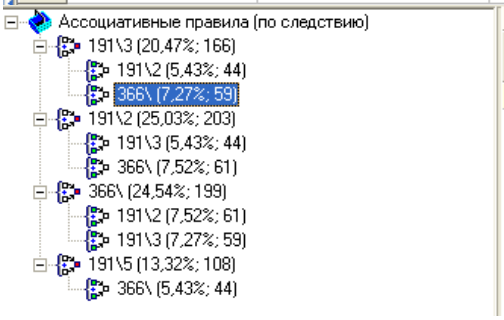
Количество правил: 3; Следствие: 366\

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	33	2,13	20,90	3,056
191\3	24	1,55	24,70	3,62
364\	26	1,68	27,10	3,963

Фінансова

Правил: 7 из 7 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	191\3	44	5,43	21,67	1,059
2	2	191\3	191\2	44	5,43	26,51	1,059
3	3	191\2	366\	61	7,52	30,05	1,225
4	4	366\	191\2	61	7,52	30,65	1,225
5	5	191\3	366\	59	7,27	35,54	1,448
6	6	366\	191\3	59	7,27	29,65	1,448
7	7	366\	191\5	44	5,43	22,11	1,660



Правило №6; Следствие: 191\3

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
366\	59	7,27	29,60	1,448

Продовження додатку В

Приватний підприємець

Правил: 7 из 7 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\3	191\2	45	2,74	21,53	0,923
2	2	191\2	366\	144	8,77	37,60	1,413
3	3	366\	191\2	144	8,77	32,95	1,413
4	4	191\3	366\	55	3,35	26,32	0,989
5	5	191\5	366\	67	4,08	38,29	1,439
6	6	364\1	366\	55	3,35	41,98	1,578
7	7	191\3 366\	191\2	21	1,28	38,18	1,637

- Ассоциативные правила (по следствию)
  - 191\2 (23,33%; 383)
    - 191\3 (2,74%; 45)
    - 366\ (8,77%; 144)
    - 191\3 И 366\ (1,28%; 21)
  - 366\ (26,61%; 437)
    - 191\2 (8,77%; 144)
    - 191\3 (3,35%; 55)
    - 191\5 (4,08%; 67)
    - 364\1 (3,35%; 55)

Количество правил: 4; Следствие: 366\

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	144	8,77	37,60	1,413
191\3	55	3,35	26,30	0,989
191\5	67	4,08	38,30	1,439
364\1	55	3,35	42,00	1,578

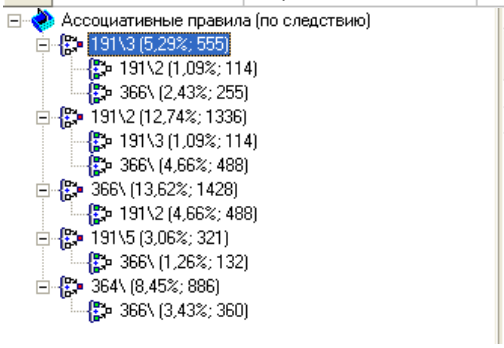
Додаток Г

Асоціативні правила та дерева правил по посадам

Керівники

Правил: 7 из 7 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	191\3	114	1,09	8,53	1,612
2	2	191\3	191\2	114	1,09	20,54	1,612
3	3	191\2	366\	488	4,66	36,53	2,681
4	4	366\	191\2	488	4,66	34,17	2,681
5	5	366\	191\3	255	2,43	17,86	3,373
6	6	366\	191\5	132	1,26	9,24	3,019
7	7	366\	364\	360	3,43	25,21	2,983



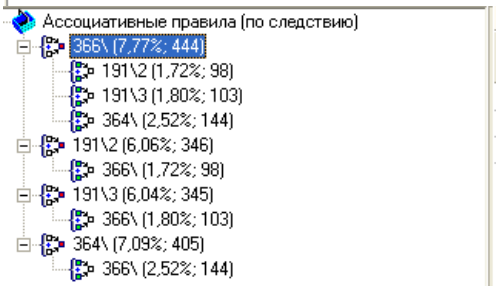
Количество правил: 2; Следствие: 191\3

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	114	1,09	8,53	1,612
366\	255	2,43	17,90	3,373

Службовець

Правил: 6 из 6 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	366\	98	1,72	28,32	3,645
2	2	366\	191\2	98	1,72	22,07	3,645
3	3	191\3	366\	103	1,80	29,86	3,842
4	4	366\	191\3	103	1,80	23,20	3,842
5	5	364\	366\	144	2,52	35,56	4,576
6	6	366\	364\	144	2,52	32,43	4,576



Количество правил: 3; Следствие: 366\

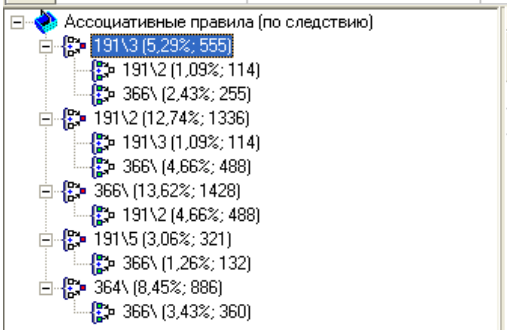
Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	98	1,72	28,30	3,645
191\3	103	1,80	29,90	3,842
364\	144	2,52	35,60	4,576

Продовження додатку Г

Депутат

Правил: 7 из 7 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	191\3	114	1,09	8,53	1,612
2	2	191\3	191\2	114	1,09	20,54	1,612
3	3	191\2	366\	488	4,66	36,53	2,681
4	4	366\	191\2	488	4,66	34,17	2,681
5	5	366\	191\3	255	2,43	17,86	3,373
6	6	366\	191\5	132	1,26	9,24	3,019
7	7	366\	364\	360	3,43	25,21	2,983



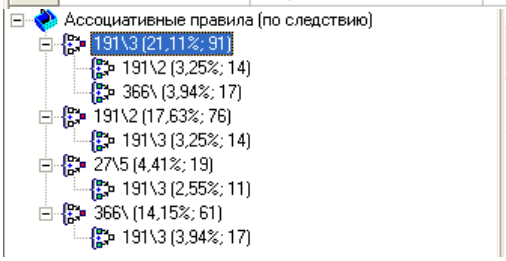
Количество правил: 2; Следствие: 191\3

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	114	1,09	8,53	1,612
366\	255	2,43	17,90	3,373

Підприємці

Правил: 5 из 5 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	191\2	191\3	14	3,25	18,42	0,872
2	2	191\3	191\2	14	3,25	15,38	0,872
3	3	191\3	27\5	11	2,55	12,09	2,742
4	4	191\3	366\	17	3,94	18,68	1,320
5	5	366\	191\3	17	3,94	27,87	1,320



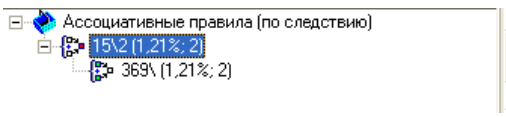
Количество правил: 2; Следствие: 191\3

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
191\2	14	3,25	18,40	0,872
366\	17	3,94	27,90	1,32

Безробітний

Правил: 1 из 1 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	369\	15\2	2	1,21	2,86	2,357



Количество правил: 1; Следствие: 15\2

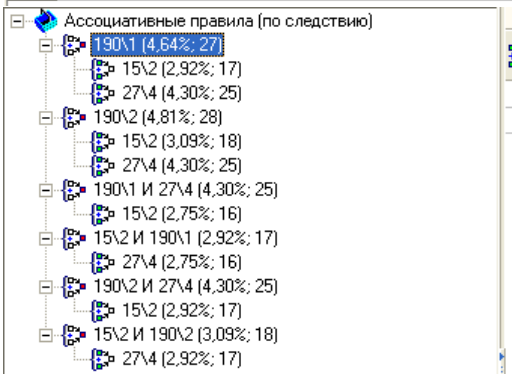
Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
369\	2	1,21	2,86	2,357

Продовження додатку Г

Інші

Правил: 8 из 8 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	15\2	190\1	17	2,92	28,81	6,211
2	2	15\2	190\2	18	3,09	30,51	6,341
3	3	27\4	190\1	25	4,30	37,88	8,165
4	4	27\4	190\2	25	4,30	37,88	7,873
5	5	15\2	190\1	16	2,75	27,12	6,313
			27\4				
6	6	27\4	15\2	16	2,75	24,24	8,299
			190\1				
7	7	15\2	190\2	17	2,92	28,81	6,708
			27\4				
8	8	27\4	15\2	17	2,92	25,76	8,328
			190\2				



Количество правил: 2; Следствие: 190\1

Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
15\2	17	2,92	28,80	6,211
27\4	25	4,30	37,90	8,165