МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

Факультет Комп'ютерних наук,
управління та адміністрування

Кафедра Інформаційних технологій

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Інструментальні засоби обробки даних в складних соціально-екологічних системах»

Виконав студент 2 курсу групи МІС- 20
спеціальності 122 Комп'ютерні науки
Обуховський Ілля Юрійович

Керівник к.т.н., доц.
Терещенко Тетяна Михайлівна

Консультант

Рецензент д.т.н., проф.
Мещеряков Володимир Іванович

Одеса 2021

ЗМІСТ

СПИСОК СКОРОЧЕНЬ І УМОВНИХ ПОЗНАЧЕНЬ

DM – Data Mining, аналіз даних.

OLAP – Online Analytical Processing, інтерактивна аналітична обробка.

SAS – система статистичного аналізу.

SQL – Structured Query Language, мова структурованих запитів.

MDX – Multidimensional Expressions, SQL-подібна мова запитів.

JHU CCSE – The Center for Systems Science and Engineering,центр системної науки та інженерії.

IR – Information Resource, інформаційний ресурс.

IT – Information Technology, інформаційні технології.

XML – Extensible Markup Language, розширювана мова розмітки.

ВСТУП

Інформаційні системи широко використовуються в різних сферах людської діяльності. Призначення таких систем полягає в зборі, збереженні та обробці інформації. Сучасні інформаційні системи ефективно працюють в економіці, екології, освіті та соціальній сфері. Все більше уваги приділяється створенню та впровадженню соціально-екологічних систем. Структурними складовими таких систем є системи управління базами даних, геоінформаційні системи та прикладні програмні продукти спеціального призначення. Останні реалізують функцію аналізу даних, пошуку залежностей, сприяють прийняттю обґрунтованих управляючих рішень. Постійне зростання обсягів даних в інформаційних системах потребує використання методів та технологій інтелектуального аналізу даних.

В сучасному світі дані стали товаром, що має високу ціну та попит. Це явище охопило багато галузей науки та виробництва. Дедалі дані стають ефективним інструментом розуміння бізнесу та його покращення в майбутньому. Великі обсяги даних та конкурентні методи обробки дозволяють отримувати необхідну інформацію та надають розуміння побудови бізнес-стратегії. Концепція інтелектуального аналізу даних отримала свій розвиток ще наприкінці минулого століття, як частина науки «інформатика». Впровадження цієї концепції для аналізу відкриття даних і знань дозволило будувати та успішно використовувати системи підтримки прийняття рішень. Особливо це важливо в складних соціально-екологічних системах тому, що саме вони впливають на якість життя окремої людини та людства загалом.

Існує велика кількість програмних інструментів для інтелектуального аналізу даних. Ці програмні продукти розповсюджуються з відкритим програмним кодом або як комерційний продукт. Такі програмні інструменти використовують в багатьох галузях виробництва та бізнесу. Крім цього, вони

широко використовуються науковцями для проведення досліджень та побудови прогнозів. Ці інструменти дозволяють достатньо легко та швидко отримати шаблони, закономірності та залежності в великих обсягах даних. Функціональні можливості такого програмного забезпечення надають користувачу інструменти для обробки як структурованих даних, та и для обробки неструктурованих даних. Більшість систем мають вбудовані алгоритми машинного навчання, а також дозволяють будувати та реалізовувати власні алгоритми ML. Використання таких алгоритмів дає можливість користувачу побачити закономірності та шаблони там, де проста аналітика не працює. Засоби візуалізації даних та контроля над робочим процесом дозволяють фахівцям ефективно аналізувати результати та оперативно приймати управлінські рішення. Ринок сучасного програмного забезпечення для інтелектуального аналізу даних пропонує різноманітні системи, які мають широкий набір функцій та інструментів для поперечної обробки даних, пошуку закономірностей та шаблонів, наочного представлення отриманих залежностей. Тому актуальною є задача дослідження такого програмного забезпечення та вибір найбільш ефективного програмного продукту, який дозволить вирішувати конкретні завдання.

Метою роботи є аналіз існуючих програмних інструментів аналізу даних та пошук найбільш ефективних для використання в соціально-екологічних інформаційних системах.

Для досягнення поставленої мети необхідно вирішити наступні завдання:

1. Провести аналіз предметної області, визначити завдання, які можливо вирішувати із використанням інтелектуальних систем аналізу даних, визначити можливість використання цих інструментів для вирішення завдань в складних соціально-екологічних системах.

2. Визначити системи та критерії для порівняльного аналізу платформ аналізу даних.

3. Провести порівняльний аналіз та визначити відповідності інструментів і задач, які можливо вирішувати з використанням таких платформ.

4. Показати практичне використання систем інтелектуального аналізу даних для прогнозування поширення епідемій.

1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1. Концепція інтелектуального аналізу даних

Data Mining (DM) – це частина науки «інформатика», яка пов'язана з дисциплінами: статистика, теорія ймовірності, штучний інтелект і машинне навчання (рис. 1) [1].

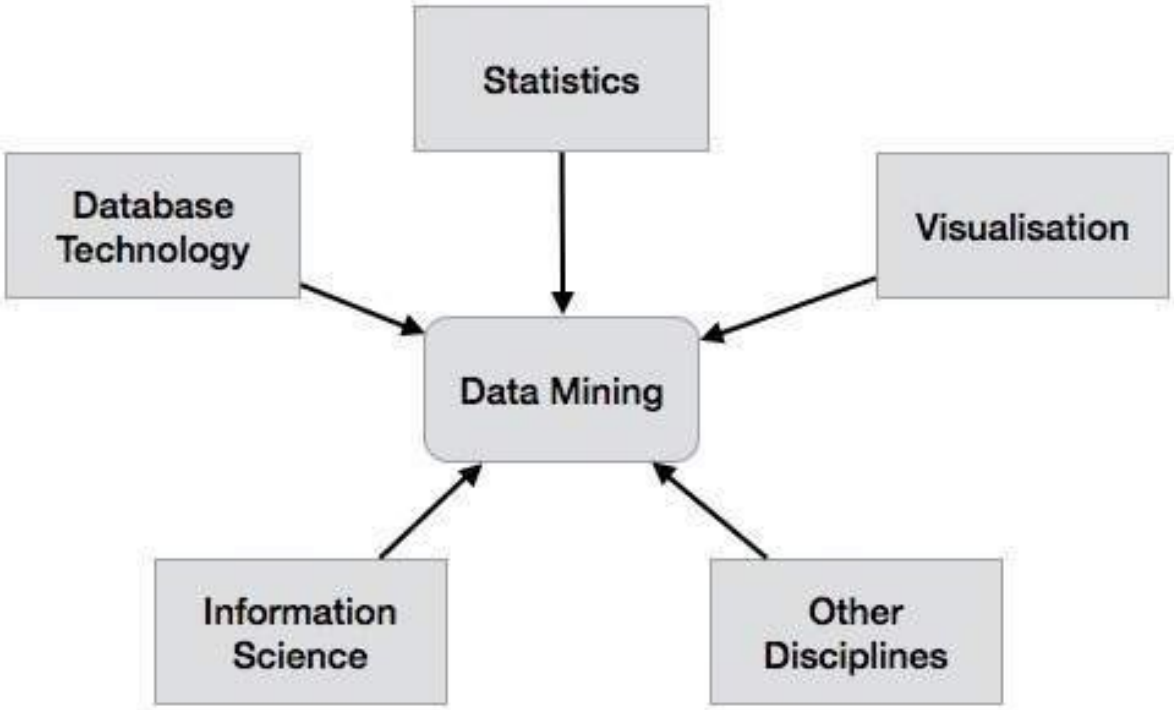

Рисунок 1 – Взаємозв’язок Data Mining з дисциплінами

Завдяки підходу до обробки даних та виявлення інформації Data Mining отримав широке використання в дослідницьких цілях. Крім того, аналіз даних дозволяє отримати набір правил, процедур і алгоритмів для створювати гіпотези, отримання шаблонів та виявлення зв'язків в величезних наборах даних. Він також включає розширений пошук, обробку та моделювання даних кількома методами одночасно. Ця сучасна сфера визнана однією з десяти найбільш впливових наук та технології з різними застосуваннями.

OLAP-системи (системи інтерактивної аналітичної обробки) надають фахівцю засоби перевірки гіпотез при аналізі даних. При цьому основним завданням аналітика є генерація гіпотез. Він вирішує її, ґрунтуючись на своїх знаннях та досвіді. Однак знання є не тільки у людини, а й у накопичених даних, які піддаються аналізу. Такі знання часто називають "прихованими", тому що вони містяться в гігабайтах та терабайтах інформації, які людина неспроможна досліджувати самостійно. У зв'язку з цим є висока ймовірність пропустити гіпотези, які можуть принести значну вигоду [1].

Вочевидь, що з виявлення прихованих знань необхідно застосовувати спеціальні методи автоматичного аналізу, з яких доводиться практично добувати знання із великих обсягів інформації. За цим напрямком міцно закріпився термін видобування даних або Data Mining.

Методи Data Mining допомагають вирішити багато завдань, із якими стикається аналітик. Серед них основними є: класифікація, регресія, пошук асоціативних правил та кластеризація.

Основні завдання аналізу даних:

1. Класифікація – це визначення класу об'єкта за його характеристиками. Слід зазначити, що у цьому завданні безліч класів, до яких може бути віднесений об'єкт, відома заздалегідь.

2. Регресія подібна до завдання класифікації та дозволяє визначити за відомим характеристиками об'єкта значення деякого його параметра. На відміну від завдання класифікації значенням параметра є не кінцева множина класів, а безліч дійсних чисел.

3. Визначення частих залежностей (або асоціацій) між об'єктами чи подіями при пошуку асоціативних правил. Знайдені залежності представляються як правила і можуть використовуватися для кращого розуміння природи даних, що аналізуються, так і для передбачення появи подій.

4. Кластеризація полягає в пошуку незалежних груп (кластерів) та їх характеристик у всій кількості даних, що аналізуються. Вирішення цього

завдання допомагає краще зрозуміти дані. Крім того, угруповання однорідних об'єктів дозволяє скоротити їх число, а отже, полегшити аналіз.

Перелічені завдання за призначенням поділяються на описові та передбачувані. Описові (descriptive) завдання приділяють увагу покращенню розуміння даних, що аналізуються. Ключовий момент у таких моделях – легкість та прозорість результатів сприйняття людиною. Можливо, виявлені закономірності будуть специфічною рисою саме конкретних даних і більше ніде не зустрінуться, але це все одно може бути корисно і тому має бути відомо. До такого виду завдань належать кластеризація та пошук асоціативних правил.

Вирішення передбачуваних (predictive) завдань розбивається на два етапи. На першому етапі на підставі набору даних з відомими результатами будується модель. На другому етапі вона використовується для прогнозування результатів на основі нових наборів даних. При цьому, потрібно, щоб побудовані моделі працювали максимально точно. До цього виду завдань відносять завдання класифікації та регресії. Сюди можна зарахувати і завдання пошуку асоціативних правил, якщо результати його вирішення можуть бути використані для передбачення появи деяких подій.

За способами вирішення завдання поділяють на supervised learning (навчання з вчителем) та unsupervised learning (навчання без вчителя). Така назва походить від терміна Machine Learning (машинне навчання), що часто використовується в англомовній літературі та позначає всі технології Data Mining.

1.2. Математичний апарат інтелектуального аналізу даних

Основою систем Data Mining є виявлення різних закономірностей у даних. При цьому застосовуються такі методи:

1. Дерева рішень.

2. Алгоритми кластеризації.
3. Регресійний аналіз.
4. Нейронні мережі.
5. Часові ряди.

Дерева рішеннь є одним з найпопулярніших підходів до вирішення задач Data Mining (рис. 2). Вони створюють ієрархічну структуру правил класифікації типу «якщо... то...», що має вигляд дерева. Для того, щоб вирішити, до якого класу віднести певний об'єкт або ситуацію, потрібно відповісти на питання, що стоїть у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд "значення параметра А більше х". Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативний – то до лівого вузла; потім знову іде питання, пов'язане з відповідним вузлом [2].

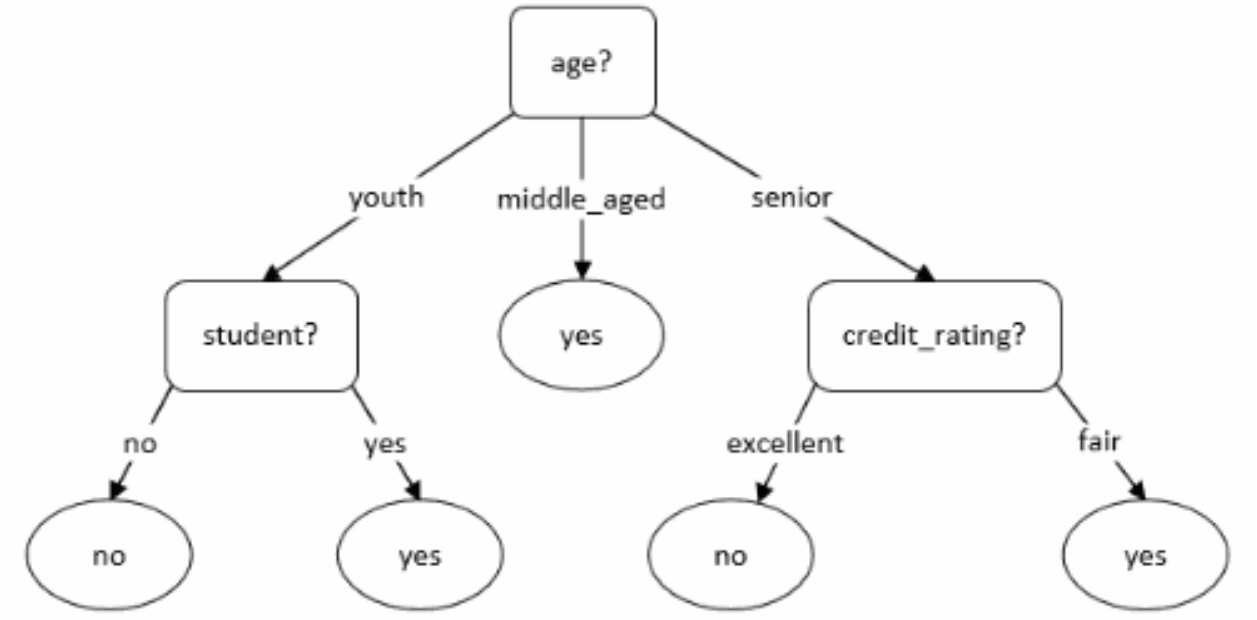

Рисунок 2 – Приклад дерева рішень

Популярність підходу пов'язана з наочністю та зрозумілістю. Але дуже гостро для дерев рішень стоїть проблема значущості. Тому, що окремим вузлам на кожному новому побудованому рівні дерева відповідає дедалі менше записів даних – дерево дробить дані на велику кількість окремих випадків. Чим більше цих окремих випадків, що менше навчальних прикладів потрапляє у кожен такий окремий випадок, тим менш впевненою стає їх класифікація. Якщо побудоване дерево надто «кущисте» – складається з невиправдано великої кількості дрібних гілочок – воно не даватиме статистично обґрунтованих відповідей. Як показує практика, у більшості систем, що використовують дерева рішень, ця проблема не знаходить задовільного рішення. Крім того, загальновідомо, і це легко показати, що дерева рішень дають корисні результати лише у разі незалежних ознак. Інакше вони лише створюють ілюзію логічного висновку [2].

Область застосування дерев рішень на даний час широка, але всі завдання, які вирішує цей апарат, можуть бути об'єднані в наступні три класи:

1. Опис даних. Дерева рішень дозволяють зберігати інформацію про дані у компактній формі, натомість можливо зберігати дерево рішень, яке містить точне опис об'єктів.

2. Класифікація. Дерева рішень добре справляються із завданнями класифікації, тобто. віднесення об'єктів до одного із заздалегідь відомих класів. Цільова змінна повинна мати дискретні значення.

3. Регресія. Якщо цільова змінна має безперервні значення, дерева рішень дозволяють встановити залежність цільової змінної від незалежних (вхідних) змінних. Наприклад, до цього класу відносяться завдання чисельного прогнозування (передбачення значень цільової змінної).

Регресійний аналіз дозволяє досліджувати форми зв'язку, що встановлюють кількісні співвідношення між випадковими величинами

досліджуваного процесу. Регресія найчастіше використовується для побудови прогнозних моделей.

Нейронні мережі є класом систем, архітектура яких намагається імітувати побудову нервової тканини з нейронів. В одній з найпоширеніших архітектур, багатошаровому персептроні (рис. 3) зі зворотним поширенням помилки, емулюється робота нейронів у складі ієрархічної мережі, де кожен нейрон більш високого рівня з'єднаний своїми входами з виходами нейронів нижчого шару. На нейрони нижнього шару подаються значення вхідних параметрів, на основі яких потрібно приймати якісь рішення, прогнозувати розвиток ситуації та ін. Ці значення розглядаються як сигнали, що передаються у вище розташований шар, послаблюючись або посилюючись залежно від числових значень (ваг, що приписуються міжнейронним зв'язкам).

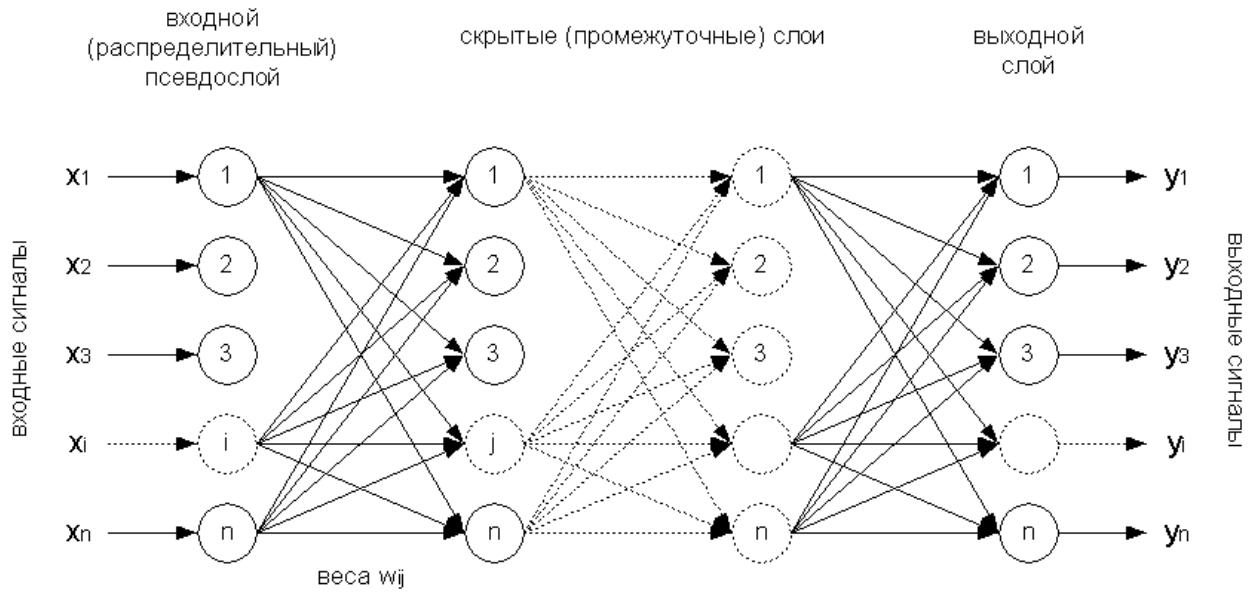

Рисунок 3 – Структура багатошарового персептрона

В результаті на виході нейрона верхнього шару виробляється деяке значення, яке розглядається як відповідь, реакція всієї мережі на введені значення вхідних параметрів. Для того щоб мережу можна було застосовувати надалі, її насамперед треба «натренувати» на отриманих

раніше даних, для яких відомі значення вхідних параметрів, і правильні відповіді на них. Це тренування полягає у підборі ваги міжнейронних зв'язків, що забезпечують найбільшу близькість відповідей мережі до відомих правильних відповідей.

Часовий ряд – це розташування в часі статистичних показників, які у своїх послідовних змінах відображають хід розвитку процесів, що вивчаються. Часові ряди досліджуються з різними цілями. В одному випадку буває достатньо отримати опис характерних рис ряду, а в іншому випадку потрібно і передбачати майбутні значення часового ряду, і управляти його поведінкою. Метод аналізу часового ряду визначається, з одного боку, цілями аналізу, з другого боку, імовірнісною природою формування його значень.

Спектральний аналіз дозволяє знаходити періодичні складові часового ряду. Кореляційний аналіз дозволяє знаходити суттєві періодичні залежності та відповідні їм затримки (лаги) як усередині одного ряду (автокореляція), так і між кількома рядами (крос-кореляція).

Моделі авторегресії та рухомого середнього орієнтовані на опис процесів, які виявляють однорідні коливання, що збуджуються випадковими впливами. Дозволяють прогнозувати майбутні значення ряду.

1.3. Області застосування технологій інтелектуального аналізу даних

Системи, що засновані на технології інтелектуального аналізу даних, використовуються в галузях та компаніях різного профілю. Однак існує ціла низка областей, для яких накопичено багатий і дуже успішний досвід застосування подібних систем.

1. Торгівля. Аналіз споживчого кошика, дослідження тимчасових шаблонів, створення прогнозуючих моделей, оптимізація складських запасів.

2. Банківська справа. Сегментація клієнтів, виявлення шахрайства із кредитними картками, прогнозування зміни клієнтури, аналіз фінансових ризиків.

3. Страховий бізнес. Сегментація клієнтів, виявлення фактів шахрайства, аналіз страхових ризиків, розробка нових товарів, розрахунок страхових премій.

4. Телекомунікації. Аналіз лояльності клієнтів, сегментування клієнтської бази та послуг, аналіз зовнішніх факторів на відмови обладнання, виявлення випадків несанкціонованого доступу до мережі.

5. Виробничі підприємства. Оптимізація закупівель, діагностика браку на ранніх стадіях, діагностика устаткування, маркетинг.

6. Нафтогазова галузь. Діагностика обладнання та нафто-газопроводів, прогнозування цін, розвідка родовищ, аналіз впливу зовнішніх та внутрішніх факторів на обсяги продажу.

Розглянемо використання інтелектуального аналізу даних в вище перелічених галузях більш докладно.

Підприємства роздрібної торгівлі збирають докладну інформацію про кожну окрему купівлю, використовуючи кредитні картки з маркою магазину та комп'ютеризовані системи контролю. Типові завдання, які можна вирішувати за допомогою технологій Data Mining у сфері роздрібної торгівлі:

1. Аналіз кошика покупця призначений виявляти товари, які покупці прагнуть купувати разом. Знання вмісту кошика необхідне для поліпшення реклами, вироблення стратегії створення запасів товарів і способів їх розкладки у торгових залах.

2. Дослідження тимчасових шаблонів допомагає торговим підприємствам приймати рішення створення товарних запасів. Воно дає відповіді на запитання типу «Якщо сьогодні покупець придбав відеокамеру, то через який час він найімовірніше купить нові батарейки та плівку?»

3. Створення прогнозуючих моделей дає можливість торговим підприємствам дізнаватися про характер потреб різних категорій клієнтів з певною поведінкою, наприклад, відомих дизайнерів, що купують товари, або відвідують розпродажі. Ці знання необхідні для розробки точно спрямованих, економічних заходів з просування товарів.

Сфера телекомунікацій характеризується високим рівенем конкуренції. Тому методи Data Mining допомагають компаніям енергійніше просувати свої програми маркетингу та ціноутворення, щоб утримати існуючих клієнтів та залучити нових. До типових заходів відносяться: аналіз записів про докладні характеристики викликів; виявлення ступеня лояльності клієнтів.

Аналіз записів про докладні характеристики дзвінків. Призначення такого аналізу – виявлення категорій клієнтів із схожими стереотипами користування їхніми послугами та розробка привабливих наборів цін та послуг.

Виявлення ступеня лояльності клієнтів. Деякі клієнти постійно змінюють провайдерів, користуючись програмами нових компаній, що стимулюють появу нових клієнтів. Data Mining використовується для визначення показників клієнтів, які, один раз скориставшись послугами цієї компанії, з великою часткою ймовірності залишаться їй вірними.

Технології Data Mining активно використовуються в центрах обробки викликів телекомунікаційних компаній.

Страхові компанії накопичують значні обсяги детальної інформації про клієнтів, про послуги, які вони використовують, суми страхових премій та виплат. Технології Data Mining дозволяють використовувати дані, що накопичилися, для вирішення наступних завдань:

1. Класифікація та кластеризація клієнтів. Система інтелектуального аналізу даних дозволяє страховій компанії проводити ефективну тарифну політику, що базується на індивідуальних уподобаннях різних категорій клієнтів.

2. Розробка нового товару. Технології Data Mining є інструментом, за допомогою якого можна спрогнозувати попит на послугу, оцінити страхові виплати та сформувати політику щодо страхових премій, що стягуються.

Більшість виробничих компаній використовують системи інтелектуального аналізу даних для вирішення наступних завдань:

1. Оптимізація логістичних ланцюжків. Data Mining дозволяє знизити витрати на логістику за рахунок ефективного прогнозування продажу товарів та закупівель сировини/комплектуючих.

2. Проведення маркетингових досліджень. Накопичені дані про обсяги збуту продукції можуть бути використані при розробці нових продуктів або підвищення ефективності рекламних кампаній.

3. Діагностика браку виробів на ранніх стадіях. Аналіз залежностей дозволяє оцінити рівень ризику виготовлення бракованого виробу на ранніх стадіях виробництва. Це дозволяє заощадити значні кошти.

Аналіз даних в соціально-екологічних системах – це окремий клас задач, які вирішуються за допомогою програмного забезпечення інтелектуального аналізу даних. Задачі цього класу різноманітні по складності пошуку закономірностей, обсягу даних для розробки, складності написання кодів для машинного навчання. Тому потребують ретельного дослідження та аналізу інструментів Data Mining.

1.4. Програмні інструменти інтелектуального аналізу даних

Програмне забезпечення для інтелектуального аналізу даних використовується комерційними компаніями та науковими товариствами для аналізу даних з широкого діапазону баз даних та виявлення закономірностей в цих великих обсягах даних. Основне призначення інструментів Data Mining – пошук даних, їхнє вилучення, поширення та монетизація інформації.

Аналіз має назву інтелектуальний тому, що може працювати як зі структурованими, так і неструктурованими даним. Основний підхід або алгоритм в таких системах – це машинне навчання. Цей підхід дозволяє визначити приховані кореляції, закономірності та тенденції, а також вказати на них користувачу [3].

Основні функціональні вимоги до програмного забезпечення Data Mining:

1. Інструменти повинні мати простий у використанні графічний інтерфейс. Ця вимога стоїть на першому місці, бо працюють з такими системами не тільки фахівці галузі комп'ютерних наук, а і спеціалісти із інших областей.

2. Якісна попередня обробка даних – це процес перетворення вихідних даних в зрозумілий формат. Під обробкою слід розуміти очищення даних, вилучення невідповідних даних, доповнення відсутніх.

3. Програмне забезпечення для аналізу даних повинно мати інструменти масштабованої обробки. Це означає, що інструменти обробки даних повинні масштабуватися відповідно кількості користувачів і розміру даних.

4. Наочне узагальнення даних. Інструменти для аналізу даних повинні мати можливість стискати данні в інформативне представлення.

1.5. Постановка задачі

Проведений аналіз предметної області дозволяє визначити задачі, які необхідно вирішити в ході дослідження, а саме:

1. Визначити програмні платформ, функціональні можливості яких необхідно дослідити.

2. Визначити критерії інструментальних засобів, за якими буде проводитися порівняльний аналіз програмних платформ DM.

3. Провести порівняльний аналіз та визначити відповідності інструментів і задач, які можливо вирішувати з використанням таких платформ.

4. Показати практичне використання систем інтелектуального аналізу даних на прикладі задачі прогнозування процесу поширення епідемій.

2. ДОСЛІДЖЕННЯ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ ОБРОБКИ ДАНИХ

2.1. Вибір інструментальних засобів для аналізу

В першому розділі була проаналізована предметна область дослідження. Основна увага біла приділена областям застосування інструментальних засобів Data Mining. Оскільки задачі інтелектуального аналізу вирішуються в різних галузях науки та виробництва, то і на ринку сучасного програмного забезпечення представлена велика кількість інструментів для аналізу даних. Вони мають як відкриту ліцензію, так і розповсюджуються на комерційній основі. Тому вибір систем для аналізу досить складна задача, оскільки на остаточний висновок щодо ефективності використанні тієї чи іншої системи впливає як характер самого завдання, так і матеріальні та технічні можливості дослідників. Враховуюче вище зазначене, в роботі спочатку буде проведено коротке дослідження характеристик програмних інструментів аналізу даних, а потім більш докладний аналіз буде виконано для декількох систем.

Інструментальні засоби Data Mining призначені для пошуку прихованих, дійсних та всіх можливих корисних шаблонів в наборах даних великого розміру. Data Mining – це метод, який допомагає досліднику виявляти несподівані/невиявлені зв'язки між даними для отримання прибутку [4].

Існує багато корисних інструментів, доступних для інтелектуального аналізу даних. Нижче наведено список найпопулярніших та найкращих програм для Data Mining з коротко описаними функціями та можливостями. Програмні інструменти не розділяються на комерційні та вільного поширення, розглядаються тільки можливості та реалізовані інструменти.

SAS Data mining – система статистичного аналізу, розробник продукту SAS. Система була розроблена для аналітики та управління даними. Вона пропонує графічний інтерфейс для нетехнічних користувачів, що важливо для областей застосування, які непов’язані з технічними та комп’ютерними науками, наприклад, медицина, торгівля тощо (рис. 4).

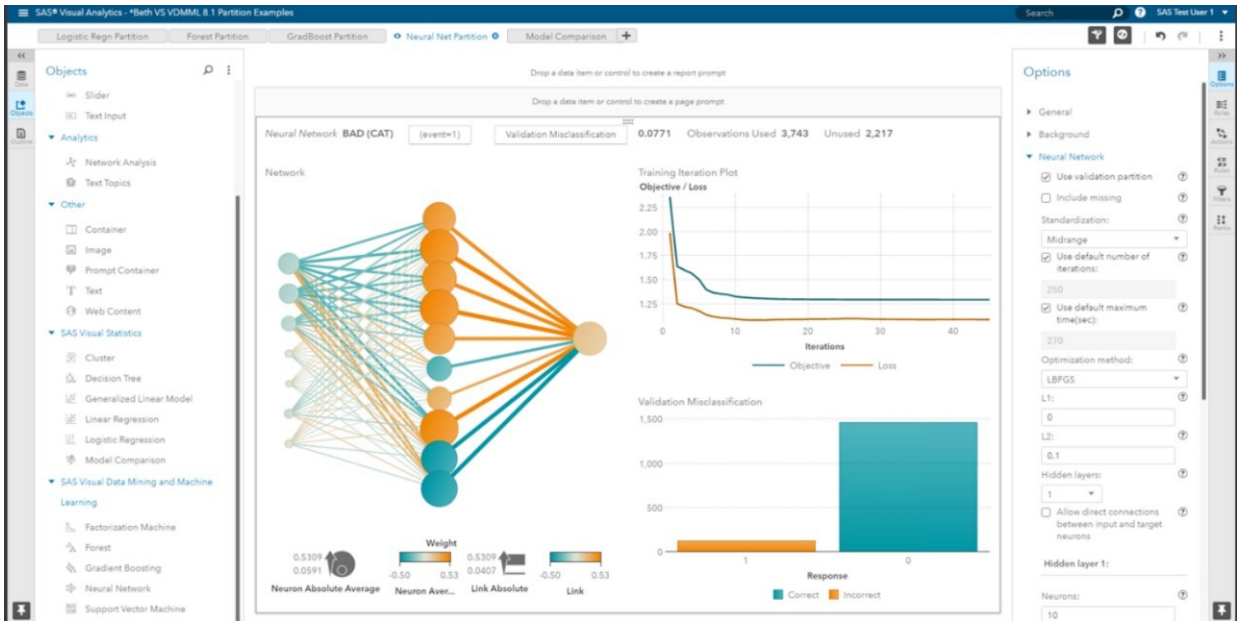

Рисунок 4 – Зовнішній вигляд інтерфейсу візуального аналізу даних

Інструменти SAS Data mining призначені для інтелектуального аналізу великих даних, аналізу тексту та оптимізації. SAS пропонує розподілену архітектуру обробки пам'яті, яка легко та ефективно масштабується.

Teradata – це масивна паралельна відкрита система обробки та розробки великомасштабних додатків сховищ даних (рис. 5). Teradata може працювати на серверній платформі Unix/Linux/Windows. Оптимізатор Teradata може обробляти до 64 з'єднань у запиті. Дані Tera мають найнижчу загальну вартість володіння. Це легко налаштувати, підтримувати та адмініструвати. Він підтримує SQL взаємодії з даними так само, які в таблицях. Це забезпечує його розширення та допомагає автоматично

розподіляти дані на диски без ручного втручання. Teradata надає утиліти завантаження та вивантаження для переміщення даних в/з системи Teradata.

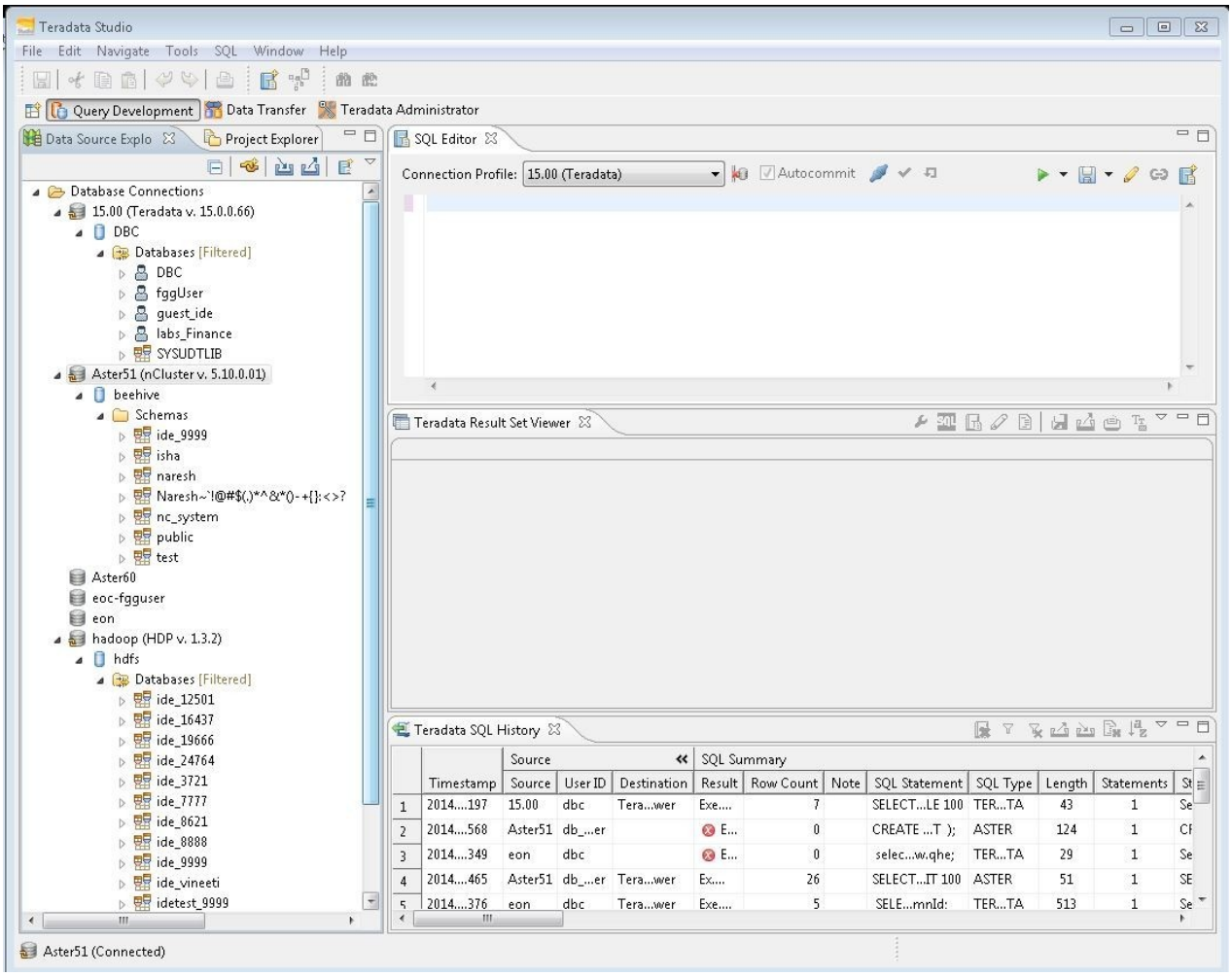

Рисунок 5 – Зовнішній вигляд інтерфейсу Teradata Studio

R-програмування – це мова для статистичних обчислень та графіки. Вона також використовується для аналізу великих даних. R пропонує широкий спектр статистичних тестів (рис. 5).

Має ефективний засіб обробки та зберігання даних, надає набір операторів для розрахунків на масивах, зокрема матрицях, забезпечує цілісний інтегрований набір інструментів для великих даних для аналізу даних, надає графічні засоби для аналізу даних, які відображаються на екрані або у друкованому вигляді.

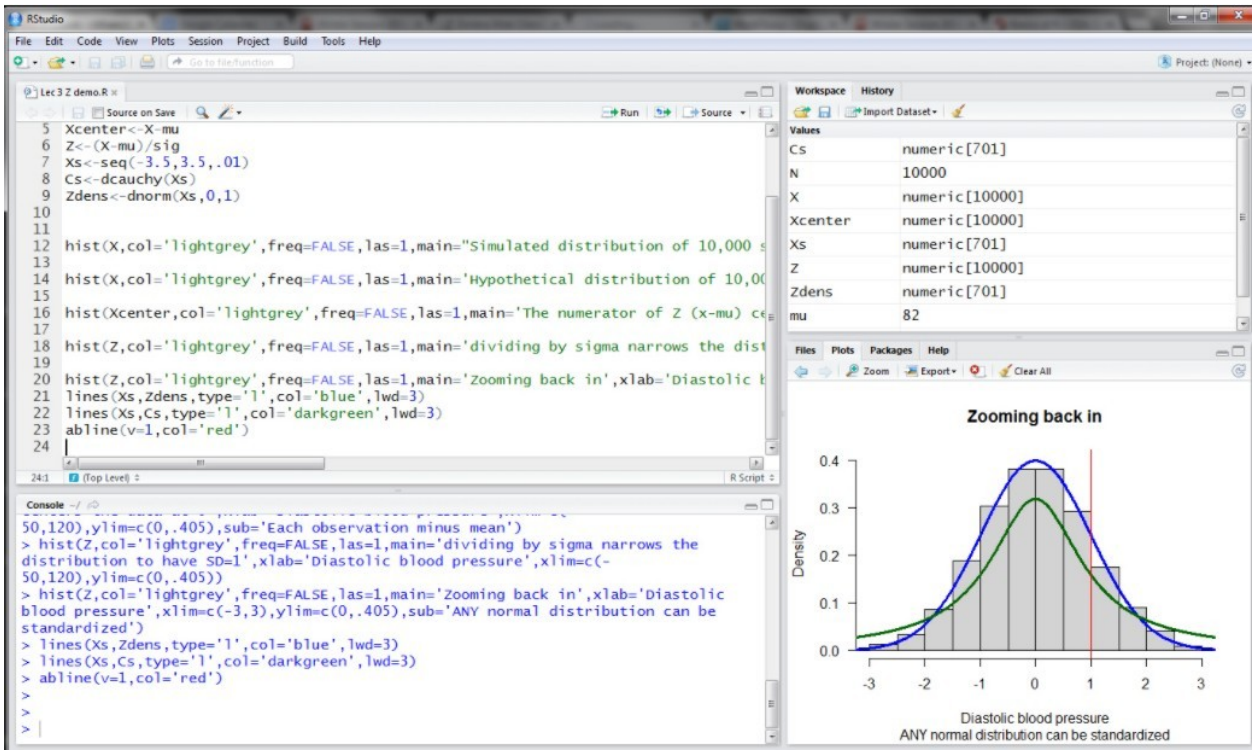

Рисунок 5 – Приклад програми на мові R з результатами роботи в середовищі RStudio [Зачем нужен язык R https://thecode.media/rrrrr/]

Python – згідно з загальним визначенням є високорівневою мовою програмування загального призначення, яка орієнтована на підвищення продуктивності та читання коду. За роки існування «пітон» обзавівся безліччю спеціалізованих бібліотек. Найбільший інтерес в питаннях аналізу даних представляють наступні:

1. Pandas – відповідає за обробку даних.

2. Numphy – працює із матрицями та багатовимірними масивами даних (рис. 6) [5].

```
In [31]: coefs = np.polyfit(x,y,1)
         predict = np.poly1d(coefs)

In [32]: x_test = np.linspace(0,16)
         y_pred = predict(x_test[:,None])
         plt.scatter(x,y)
         plt.plot(x_test,y_pred,c='r')
         plt.show()
```

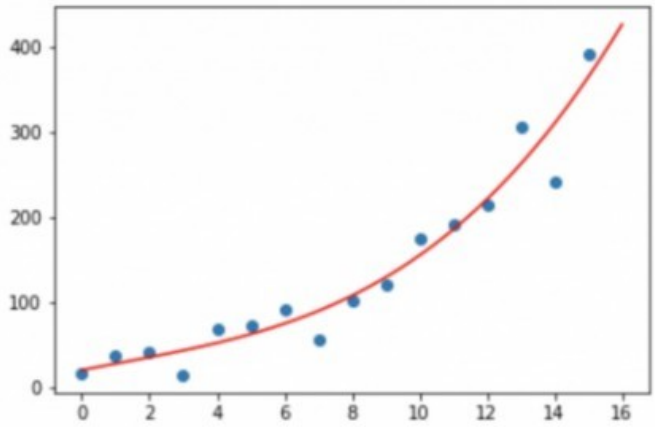

Рисунок 6 – Приклад прогнозування з використанням функції polyfit

3. Statsmodels – містить основні статистичні функції та моделі.

4. Sklearni Pybrain – спеціалізуються на алгоритмах машинного навчання.

5. Matplotlib – відповідає за візуалізацію результатів аналізу та моделювання.

Крім добре документованих бібліотек, Python відрізняється гнучкістю та зрозумілим синтаксисом – завдяки останньому він приємний у роботі. Важливо і те, що Python має величезну спільноту відданих «фанатів», справжніх фахівців своєї справи, тому мова не перестає розвиватися.

Board – це інструментарій управління розвідкою. Він поєднує у собі функції бізнес-аналітики та корпоративного управління ефективністю. Він призначений для надання бізнес-аналітики та бізнес-аналітики в одному пакеті. Дозволяє аналізувати, моделювати, планувати та прогнозувати, використовуючи єдину платформу (рис. 7).

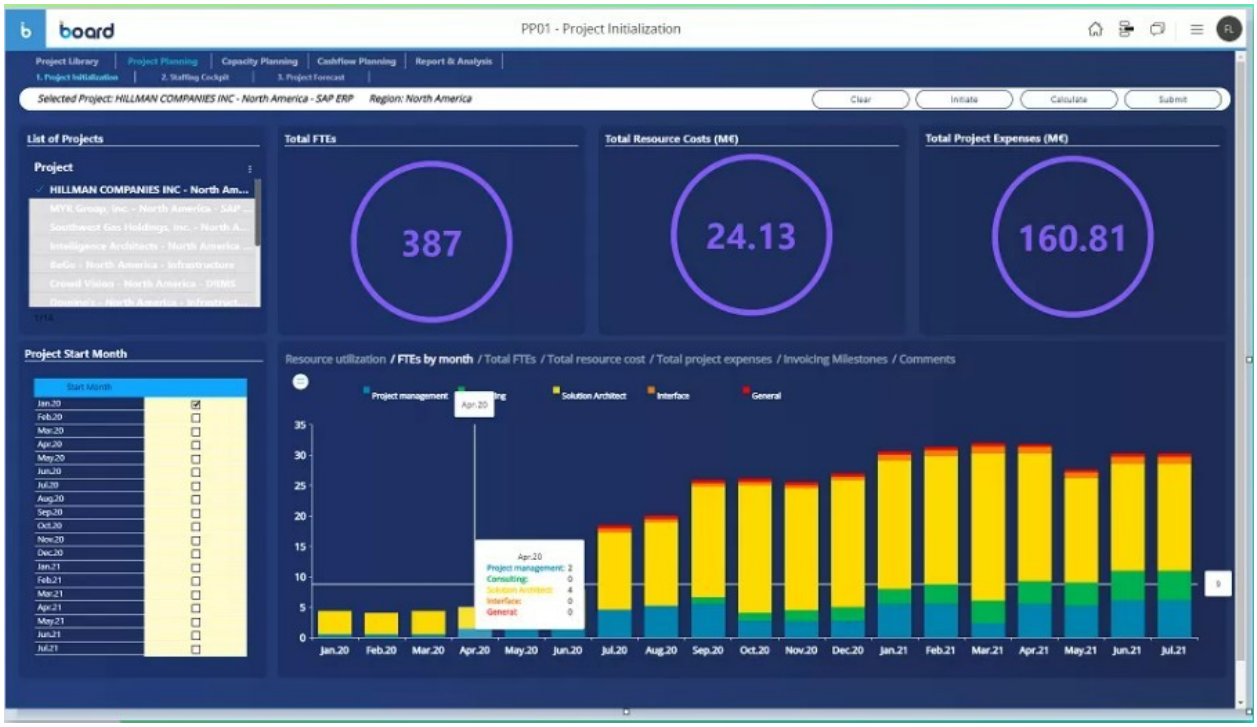

Рисунок 7 – Графічний інтерфейс платформи Board

Має функції створення індивідуальних аналітичних та планових програм. Власна платформа допомагає складати звіти, одержуючи доступ до кількох джерел даних.

Dundas – це готовий до роботи інструмент для збору даних, який можна використовувати для створення та перегляду інтерактивних інформаційних панелей, звітів тощо.

Dundas BI розгортається як центральний портал даних для організації. Серверний додаток має повну функціональність продукту, передбачена інтеграція та доступ до всіх видів джерел даних. Структурна схема порталу представлена на рис. 8.

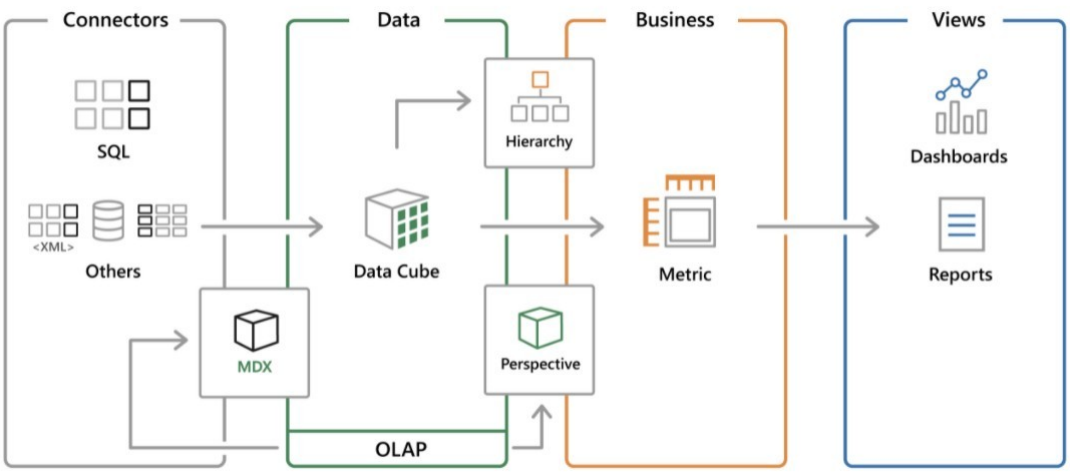

Рисунок 8 – Структурна схема порталу Dundas

Візуалізація даних налаштовується під потреби конкретного завдання та можливості користувачів, реалізована можливість візуалізувати дані через карти.

Inetsoft – це корисна платформа для інтелектуального аналізу даних, яка дозволяє швидко та гнучко перетворювати дані з різних джерел (рис. 9). Це дає можливість отримати доступ до структурованих та напівструктурованих джерел, локальних додатків.

Платформа дозволяє оптимізувати програми для споживання та оновлення даних. Передбачена функція масштабування для великих масивів даних користувачів за допомогою платформи Inbuilt Spark. Створення розбитих на сторінках звітів з вбудованою бізнес-логікою та параметризацією реалізовано простими та доступними інструментами.

Рисунок 9 – Графічний інтерфейс платформи Inetsoft

H_3O – інструмент для аналізу даних із відкритим вихідним кодом. Він використовується для аналізу даних, що зберігаються у прикладних системах хмарних обчислень. Дозволяє використовувати переваги обчислювальної потужності розподілених систем та обчислень у пам'яті, швидко та легко впроваджувати у виробництво Java та двійковий формат.

Weka – це ціла колекція інструментів та алгоритмів для аналізу даних та прогнозування. Має зручний інтерфейс (наприклад, текстовий рядок для введення команд); реалізовано перетворення даних (зокрема попередня обробка «сирих» даних); передбачена підтримка безлічі алгоритмів машинного навчання та можливість їх швидкого застосування. Характеризується зручним виведенням результатів роботи алгоритму (легко порівнювати точність різних моделей) (рис. 10).

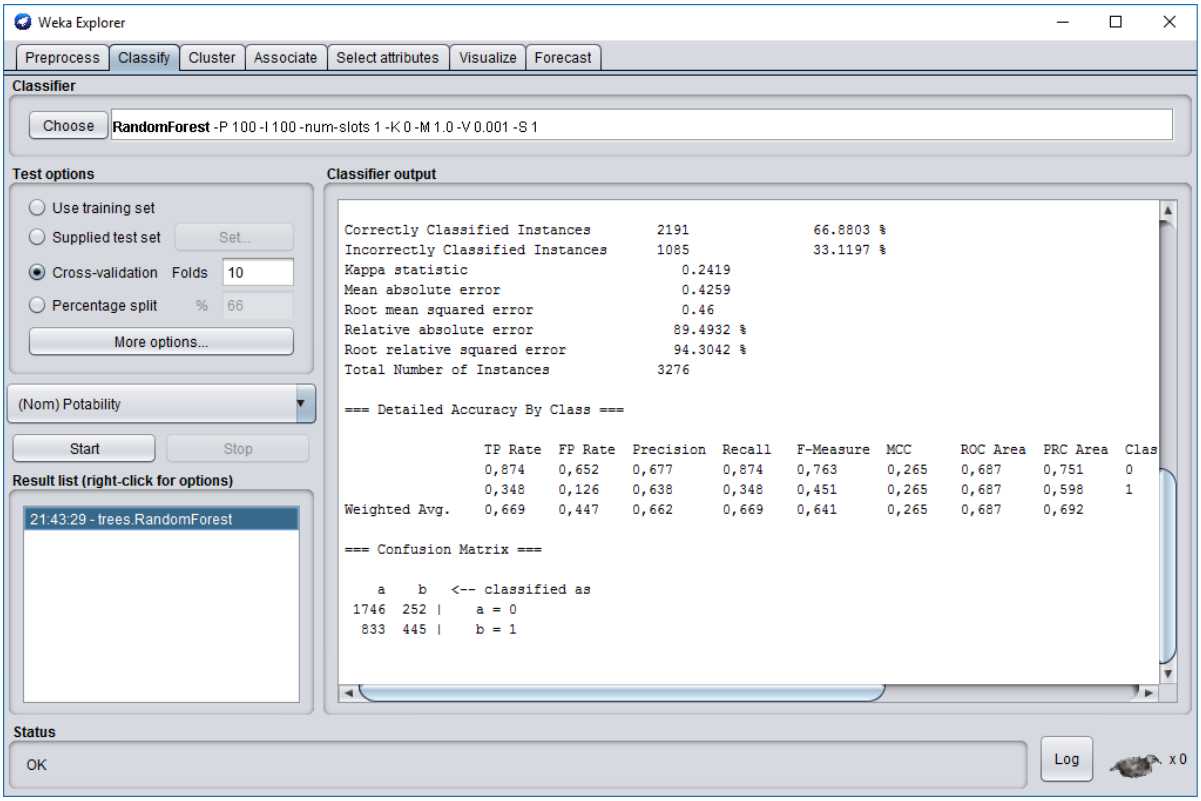

Рисунок 10 – Результати навчання моделі в Weka

Передбачено вибір ознак та візуалізація даних; реалізована можливість проведення експериментів (можна запускати відразу кілька алгоритмів на різних завданнях та отримати загальний звіт). Має можливість представлення всього процесу розв'язання задачі у формі графа.

KNIME Analytics Platform – це програмне забезпечення з відкритим кодом для створення науки про дані. Має інтуїтивно зрозумілий інтерфейс, який відкритий та постійно інтегрує нові розробки. KNIME використовується для розуміння даних, проектування робочих процесів обробки даних та створення компонентів, які можливо повторно використовувати (рис. 11) [6].

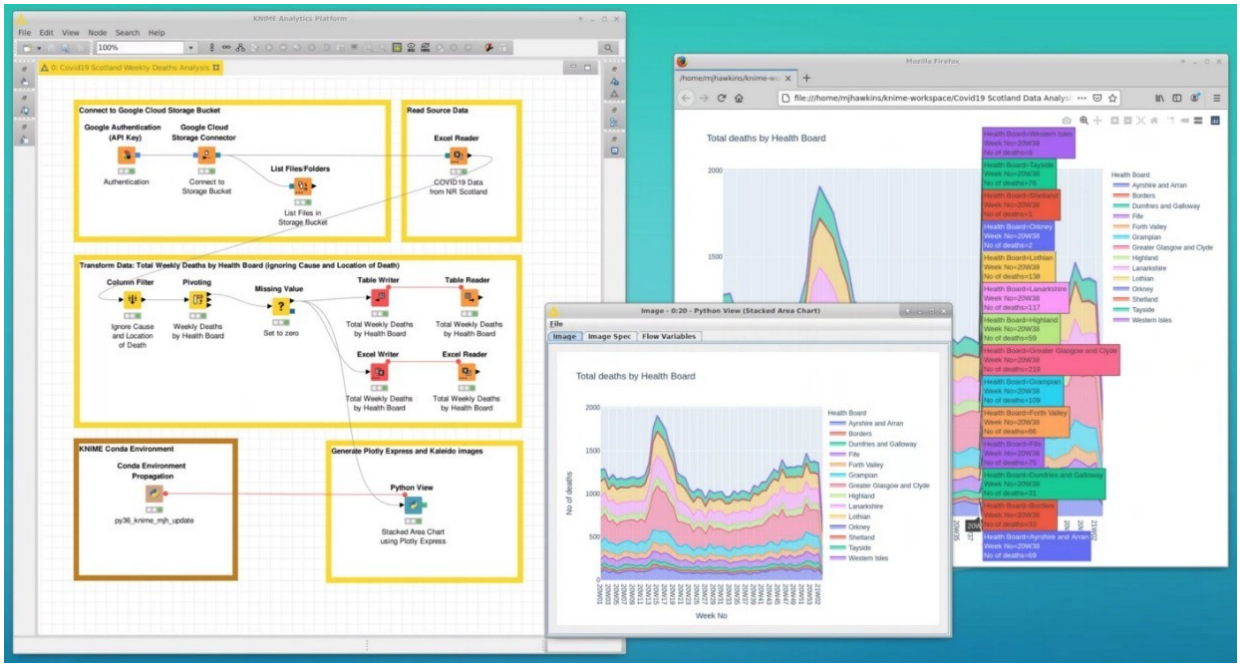

Рисунок 11 – Приклад робочого процесу платформи KNIME Analytics

Створює візуальні робочі процеси для аналізу даних за допомогою інтуїтивно зрозумілого графічного інтерфейсу в стилі перетягування без кодування. Об'єднує інструменти з різних доменів з власними вузлами KNIME в одному робочому процесі, включаючи створення сценаріїв в R&Python, машинне навчання або конектори з Apache Spark.

RapidMiner підтримує безліч стандартних завдань, що стосуються перетворення даних, статистики, машинного навчання та візуалізації. Весь процес аналізу даних представляється у вигляді інтерактивного графа – послідовності операторів, при цьому користувачу доступні оператори Weka і R (рис. 12).

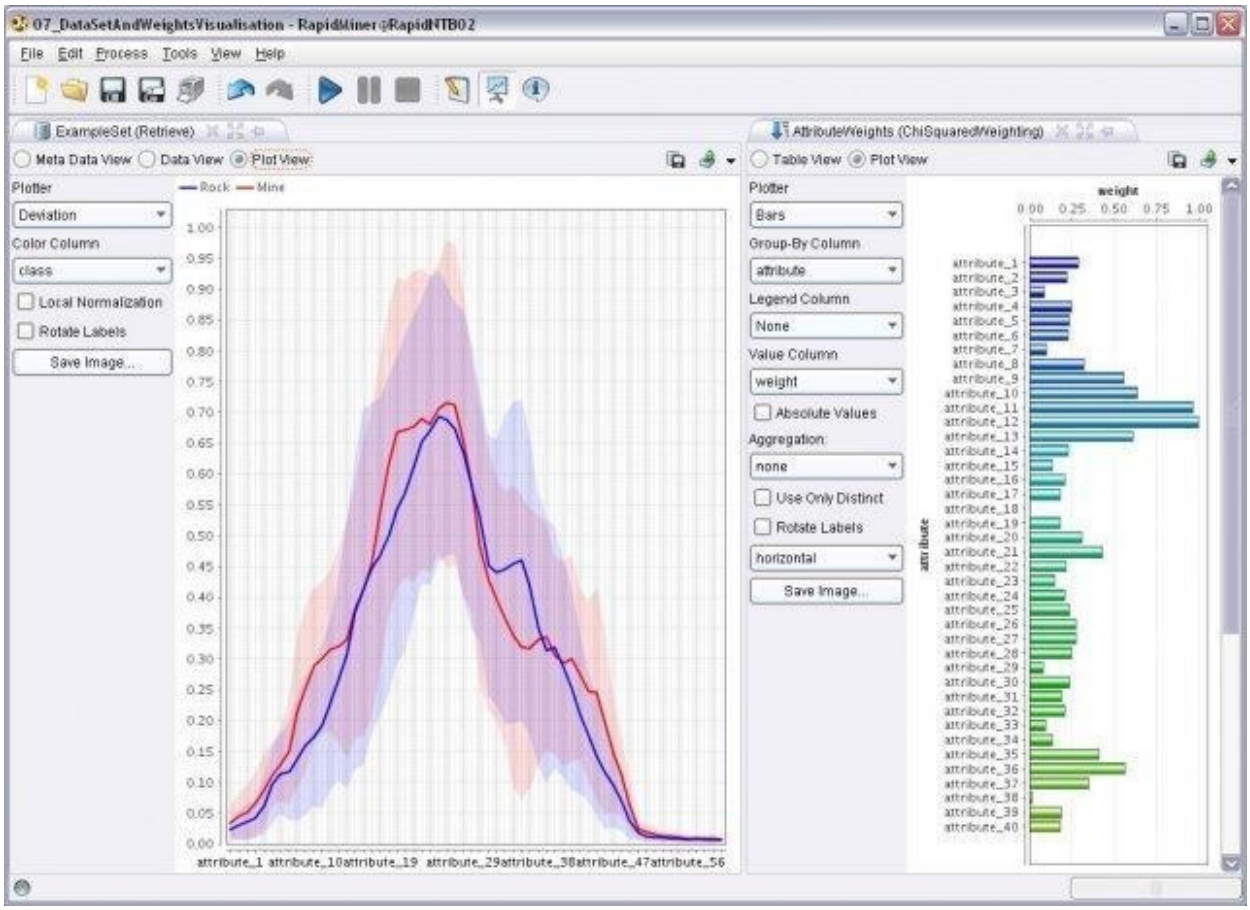

Рисунок 12 – Приклад робочого процесу в RapidMiner

RapidMiner реалізує достатню кількість методів управління даними: завантаження даних, перетворення даних, моделювання даних та методи візуалізації даних. Працює з кількома джерелами даних: Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, текстові файли та ін. Розроблені та впроваджені нові шаблони: скорочення відтоку, аналіз настроїв, профілактичне обслуговування та прямий маркетинг. Працює на всіх основних платформах та операційних системах.

RapidMiner Radoop може підключатися до багатьох кластерів Hadoop: Cloudera Distribution, включаючи Apache Hadoop (CDH), Hortonworks Data Platform (HDP), Apache Hadoop з Hive, Amazon Elastic MapReduce, MapR Hadoop і DataStax Enterprise. Зберігає потокові дані та результати аналітики у

численних базах даних, включаючи Cassandra, MongoDB, Redis, Apache Solr та ін.

Слід звернути увагу на те, що задачі соціально-екологічного аналізу даних відрізняються різноманітністю та складністю. Тому при виборі інструментальних засобів необхідно враховувати не тільки функціональні можливості системи, а і доступність інструкцій користувача та керівництв розробника як від виробника, так і від фахівців, які використовують інструменти аналізу для вирішення своїх задач.

З урахуванням віще зазначеного, для подальшого детального аналізу інструментів було обрано п’ять платформ, а саме, Python, R, Weka, Knime, RapidMiner, як найбільш задовольняючі висунутим умовам.

2.2. Вибір критеріїв для аналізу

При виборі критеріїв для аналізу платформ слід враховувати характер задач, які будуть вирішуватися з використанням програмних інструментів аналізу даних. Оскільки всі задачі соціально-екологічного характеру пов’язані з обробкою великих обсягів даних, у якості першого критерію було обрано «Якість обробки даних». Цей критерій включає в себе оцінку необхідних практичних навичок фахівця, можливість робити складні перетворення даних, якість простих перетворень даних, наприклад, нормалізація та ін.

Оцінка кожного критерія була зроблена за шкалою від 1 до 4, де 1 – слабо відповідає, 4 – відповідає повністю. Результати оцінки кожної з обраних платформ за визначеними критеріями та загальна оцінка по критерію «Якість обробки даних» наведені в таблиці 1.

Таблиця 1 – Результати оцінки платформ по критерію «Якість обробки даних»

Якість обробки даних	Python	R	Weka	Knime	Rapid Miner
Практичні навички	4	2	1	1	1
Можливість робити складні перетворення даних	3	3	1	1	1
Якість простих перетворень даних	4	4	4	4	4
Загальна оцінка	11	9	6	6	6

Наступний критерій «Візуалізація даних» складається з двох оцінок гнучкість та естетичний вигляд. Мета візуалізації – уявити спрощено інформацію, дозволити користувачу з одного погляду скласти про неї певну думку. При цьому, ідеальна візуалізація є зрозумілою сама по собі, і може зберегти свій сенс, навіть втративши супроводжуючий текст. Користувачі, що приймають рішення на високому рівні, як правило, не мають часу вникати в нескінченні ряди даних, тому їм потрібен матеріал, на підставі якого вони можуть швидко прийняти якісне рішення та оцінити ситуацію, не заглиблюючись в аналіз первинної інформації. Згідно з дослідженнями людського сприйняття, найбільшу частину інформації надходить до нас через очі. Саме тому якість візуалізації є вкрай важливим елементом прийняття рішень. Оцінка кожного критерія була зроблена за шкалою від 1 до 4, де 1 – слабо відповідає, 4 – відповідає повністю. Результати оцінки кожної з обраних платформ за визначеними критеріями та загальна оцінка по критерію «Візуалізація даних» наведені в таблиці 2.

Таблиця 2 – Результати оцінки платформ по критерію «Візуалізація даних»

Візуалізація даних	Python	R	Weka	Knime	Rapid Miner
Гнучкість процесу візуалізації	4	4	2	2	2
Естетичний вигляд результатів (наочність, прозорість, зрозумілість)	2	4	2	2	2
Загальна оцінка	6	8	4	4	4

Наступний критерій аналізу «Машинне навчання» складається з оцінки кількості методів навчання та гнучкості налаштування параметрів. Алгоритми машинного навчання дають можливість користувачам досліджувати та аналізувати складні набори даних та знаходити в них залежності. В моделі машинного навчання встановлюються або виявляються закономірності, за допомогою яких користувачі можуть створювати прогнози або класифікувати інформацію. В алгоритмах машинного навчання використовуються параметри, що базуються на навчальних даних (підмножина даних, що представляє ширший набір). При розширенні навчальних даних для більш реалістичного уявлення світу за допомогою алгоритму обчислюються більш точні результати. В різних алгоритмах застосовуються різні методи аналізу даних. Вони часто групуються за методами машинного навчання, у межах яких використовуються: контрольоване навчання, неконтрольоване навчання та навчання з підкріпленням. У найбільш популярних алгоритмах для прогнозування цільових категорій, пошуку незвичайних точок даних, прогнозування значень та виявлення подібності використовуються регресія та класифікація.

Оцінка кожного критерія була зроблена за шкалою від 1 до 4, де 1 – слабо відповідає, 4 – відповідає повністю. Результати оцінки кожної з обраних платформ за визначеними критеріями та загальна оцінка по критерію «Машинне навчання» наведені в таблиці 3.

Таблиця 3 – Результати оцінки платформ по критерію «Машинне навчання»

Машинне навчання	Python	R	Weka	Knime	Rapid Miner
Кількість методів навчання	4	4	4	3	3
Гнучкість налаштування параметрів	4	4	2	2	2
Загальна оцінка	8	8	6	5	5

Наступний критерій аналізу «Представлення результатів роботи» складається з оцінки гнучкості представлених результатів та витрат часу на налаштування параметрів. Цей критерій має важливе значення оскільки впливає на витрати часу, які і так значні при роботі з великими обсягами даних. Гнучкість налаштування параметрів результатів роботи є обов’язковою умовою в системах інтелектуального аналізу даних, оскільки напряму впливає на наочність отриманих результатів та зрозумілість виявлених залежностей. Крім того цей параметр впливає на коректність прийнятих управлінських рішень.

Оцінка кожного критерія була зроблена за шкалою від 1 до 4, де 1 – слабо відповідає, 4 – відповідає повністю. Результати оцінки кожної з обраних платформ за визначеними критеріями та загальна оцінка по критерію «Представлення результатів роботи» наведені в таблиці 4.

Таблиця 4 – Результати оцінки платформ по критерію «Представлення результатів роботи»

Представлення результатів роботи	Python	R	Weka	Knime	Rapid Miner
Гнучкість	4	4	1	2	2
Витрати часу на налаштування параметрів	1	2	4	3	3
Загальна оцінка	5	6	5	5	5

Наступний критерій аналізу «Швидкість отримання попередніх результатів» складається з оцінки витрат часу на написання коду та швидкості вводу результатів. Цей критерій впливає на оперативність прийняття управлінських рішень і для деяких важливих задач має критичне значення.

Оцінка кожного критерія була зроблена за шкалою від 1 до 4, де 1 – слабо відповідає, 4 – відповідає повністю. Результати оцінки кожної з обраних платформ за визначеними критеріями та загальна оцінка по критерію «Швидкість отримання попередніх результатів» наведені в таблиці 5.

Таблиця 5 – Результати оцінки платформ по критерію «Швидкість отримання попередніх результатів»

Швидкість отримання попередніх результатів	Python	R	Weka	Knime	Rapid Miner
Витрати часу на написання коду	1	1	4	4	4
Швидкость вводу результатів	2	3	4	4	4
Загальна оцінка	3	4	8	8	8

Наступний критерій аналізу «Наочність процесу аналізу даних» характеризує можливості платформи відображати процес аналізу даних в реальному часі. Сучасні платформи аналізу даних надають користувачу можливості нагляду та контролю за процесом аналізу, що робить системи більш гнучкими та ефективними.

Оцінка кожного критерія була зроблена за шкалою від 1 до 4, де 1 – слабо відповідає, 4 – відповідає повністю. Результати оцінки кожної з обраних платформ за визначеними критеріями та загальна оцінка по критерію «Наочність процесу аналізу даних» наведені в таблиці 6.

Таблиця 6 – Результати оцінки платформ по критерію «Наочність процесу аналізу даних»

Наочність процесу аналізу даних	Python	R	Weka	Knime	Rapid Miner
Наочність	1	1	2	4	4
Загальна оцінка	1	1	2	4	4

2.3. Результати аналізу програмних інструментів

В результаті проведеного аналізу обраних платформ аналізу даних за обраними критеріями отримали загальну оцінку для кожної системи, які представлені в табл. 7. Загальна оцінка для кожної системи визначалась як сума оцінок по критеріям, які були визначені в попередньому підрозділі.

Таблиця 7 – Результати проведеного порівняльного аналізу платформ за обраними критеріями

Критерій оцінки	Python	R	Weka	Knime	Rapid Miner
Якість обробки даних	11	9	6	6	6
Візуалізація даних	6	8	4	4	4
Машинне навчання	8	8	6	5	5
Представлення результатів роботи	5	6	5	5	5
Швидкість отримання попередніх результатів	3	4	8	8	8
Наочність процесу аналізу даних	1	1	2	4	4
Загальна оцінка	34	36	31	32	32

Загальні висновки, які можна зробити за результатами проведеного аналізу, наступні:

1. Максимальну кількість балів набрали Python і R, визначені загальні оцінки складають 34 та 36 відповідно. Інші системи мають приблизно однакові загальні оцінки 31-32.

2. Розбіжність оцінок по кожному критерію свідчить про те, що при виборі платформи для аналізу даних слід враховувати технологічні вимоги до програмних інструментів та характер завдання, яке необхідно вирішити за допомогою цього ПЗ. Наприклад, Python і R показали найкращі результати для задач обробки та візуалізації даних.

3. За критерієм «Швидкість отримання попередніх результатів» максимальні оцінки отримали платформи Weka, Knime, RapidMiner. Тому для задач, які потребують оперативного отримання результатів, слід використовувати саме ці платформи.

4. Приблизно однакові оцінки всі системи отримали за критерієм «Представлення результатів роботи». Тому даний критерій можна не розглядати при виборі платформи для конкретної задачі.

5. В задачах, для яких важлива наочність процесу аналізу даних, слід використовувати Knime, RapidMiner, оскільки тільки ці дві системи отримали високі оцінки за даним критерієм.

6. Для задач, основною метою яких є машинне навчання, ефективніше використовувати Python і R тому, що вони дають більше свободи для розробників в питаннях створення та реалізації стандартних або спеціалізованих алгоритмів машинного навчання.

В результаті проведеного порівняльного аналізу платформ для аналізу даних не вдалося визначити найбільш ефективну систему для вирішення складних соціально-екологічних задач. Це пов’язано з тим, що три системи отримали приблизно однакові оцінки за всіма обраними критеріями (Weka, Knime, RapidMiner).

Але можна зробити висновок, що Python і R вимагають додаткових знань та навичок програмування. Це обмежує використання цих систем в галузях, що не пов’язані напряму з комп’ютерними науками. Не зважаючи на те, що доступна велика кількість алгоритмів машинного навчання, які реалізовані цими мовами програмування.

Тому було прийнято рішення провести додаткове дослідження платформ Weka, Knime, RapidMine з метою виявлення найбільш ефективної системи. Методом дослідження було обрано експертну оцінку тому, що простий порівняльний аналіз не дав бажаних результатів.

3. РЕЗУЛЬТАТИ ЕКСПЕРТНОЇ ОЦІНКИ ФУНКЦІОНАЛЬНИХ ВЛАСТИВОСТЕЙ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ ОБРОБКИ ДАНИХ

В мережі Інтернет існують ресурси, які збирають оцінки експертів та користувачів систем аналізу даних. Найбільш повну інформацію про сучасні системи аналізу даних містить інтернет-портал www.predictiveanalyticstoday.com. На цьому ресурсі зібрана інформація про існуючи системи аналізу даних, які згруповані за видами принципу розповсюдження. Окремо наведені дані про системи з відкритим кодом та системи, які розповсюджуються з комерційною ліцензією [3].

Результати оцінки інструментів представлені окремими рейтингами для платного програмного забезпечення та продуктів з відкритим кодом (рис. 13, рис. 14).

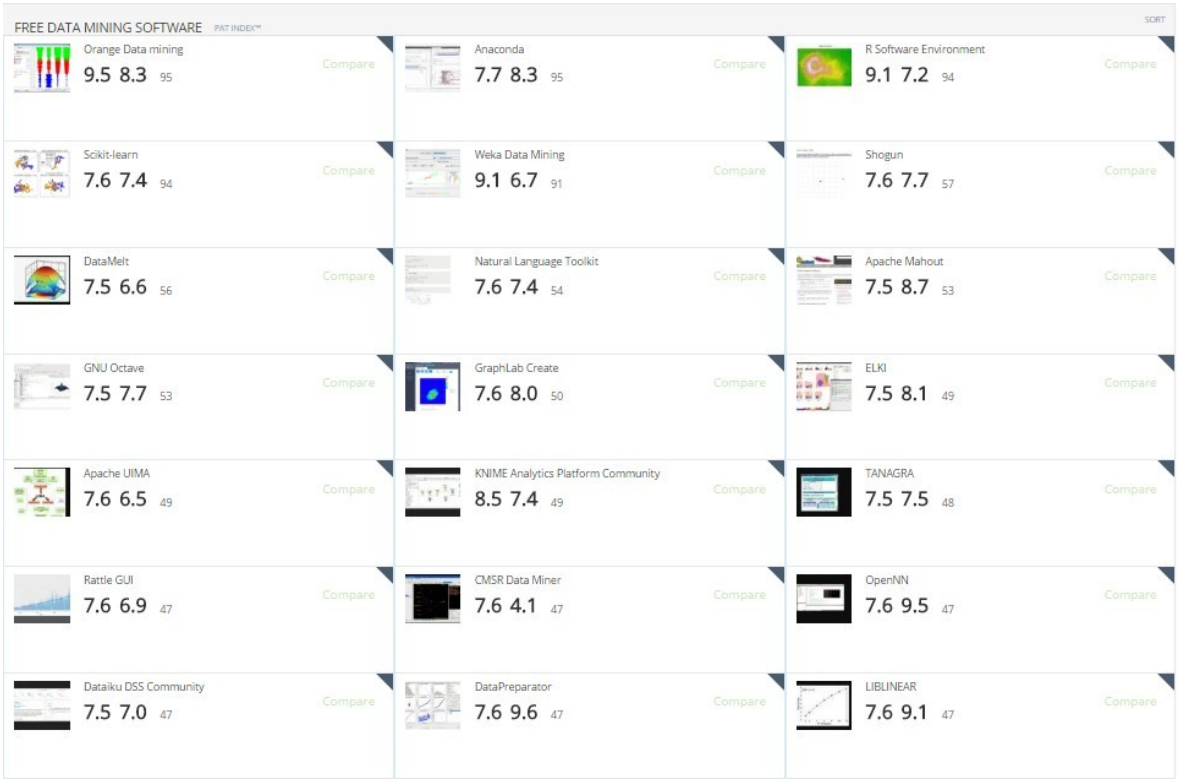

Рисунок 13 – Результати аналізу Free Data Mining Software

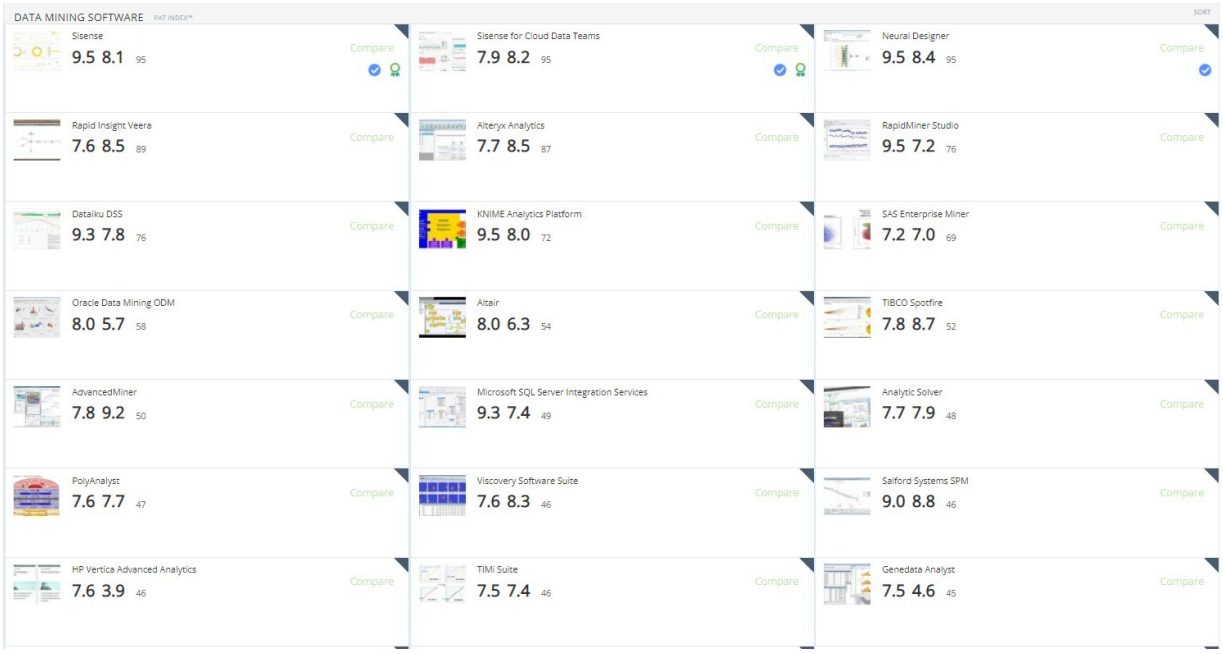

Рисунок 14 – Результати аналізу Data Mining Software

Кожен програмний продукт має оцінку редакції та оцінку користувачів. Рейтинг складається з набору властивостей програмного забезпечення: простота використання, функціональність, розширення можливостей, інтеграція тощо.

Розглянемо послідовно системи Weka, Knime, RapidMine. Оцінка експертів проводиться за наступними критеріями:

1. Простота використання.
2. Характеристики та функціональність.
3. Розширені функції.
4. Інтеграція з іншими системами.
5. Продуктивність.
6. Підтримка клієнтів.
7. Простота реалізації.
8. Оновлення та рекомендації.

3.1. Платформа Weka Data Mining

Це програмне забезпечення являє собою набір алгоритмів машинного навчання для інтелектуального аналізу даних, розповсюджується з відкритим кодом (рис. 13) [3].

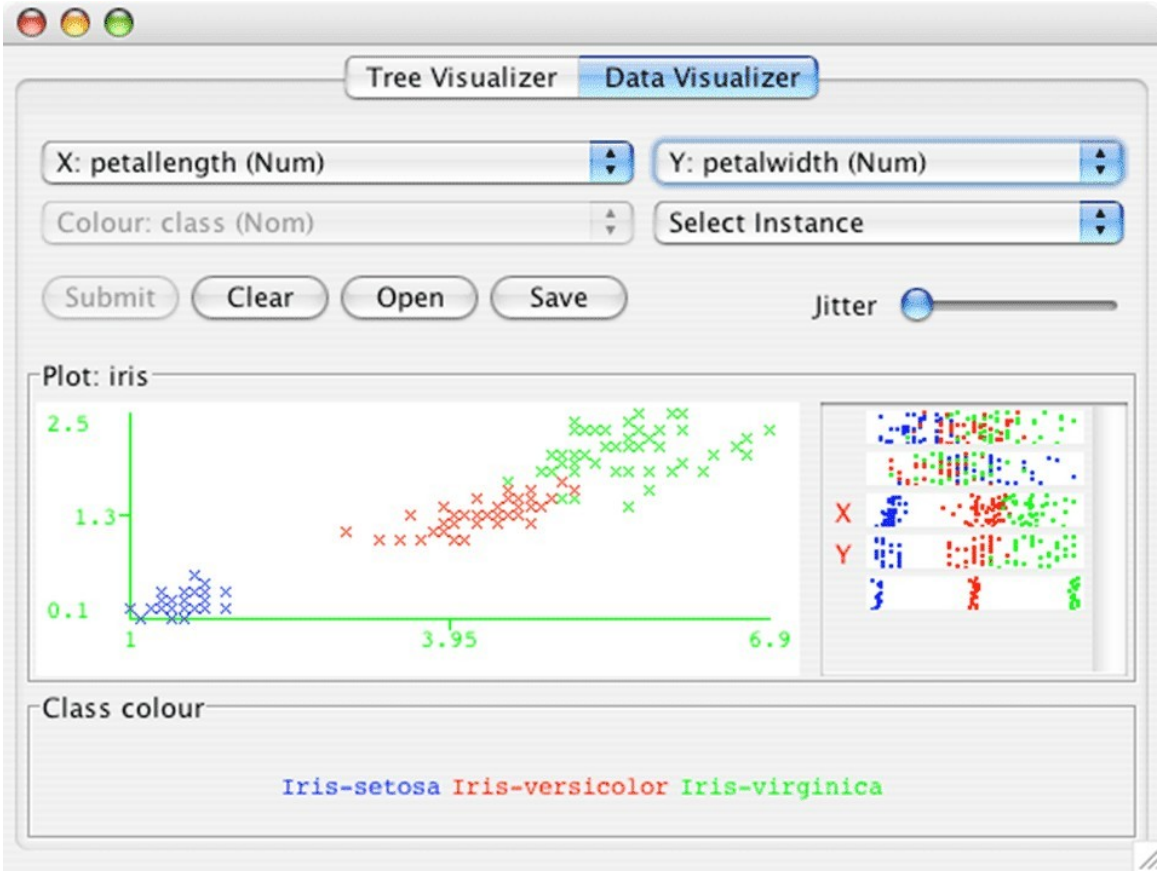

Рисунок 13 – Графічний інтерфейс Weka

Основні функції, які реалізовані в системі Weka:

- машинне навчання;
- аналіз даних;
- попередня обробка даних;
- класифікація даних;
- регресія даних;

- кластеризація даних;
- плавила асоціації;
- вибір атрибутів та налаштування;
- експерименти;
- контроль робочого процесу;
- візуалізація даних.

До переваг системи Weka Data Mining можна віднести:

- простота використання;
- портативність;
- безкоштовне розповсюдження;
- адаптивність для створення нових алгоритмів машинного навчання;
- безкоштовні онлайн-курси.

Таблиця 8 – Рейтинг Weka за обраними критеріями

Критерії оцінювання	Оцінка редактора	Оцінка користувачів (загальна)
Простота використання	9,1	6,3
Характеристики та функціональність	9,2	6,4
Розширені функції	9,2	6,8
Інтеграція з іншими системами	9,0	7,4
Продуктивність	9,1	6,3
Підтримка клієнтів	9,0	7,2
Простота реалізації		7,4
Оновлення та рекомендації		7,1
Загальний рейтинг	9,1	6,6

3.2. Платформа аналітики Knime

Knime Analytics Platform – це відкрите рішення для інноваційної обробки даних. Розповсюджується як безкоштовне програмне забезпечення для прогнозування та аналітики (рис. 14) [3].

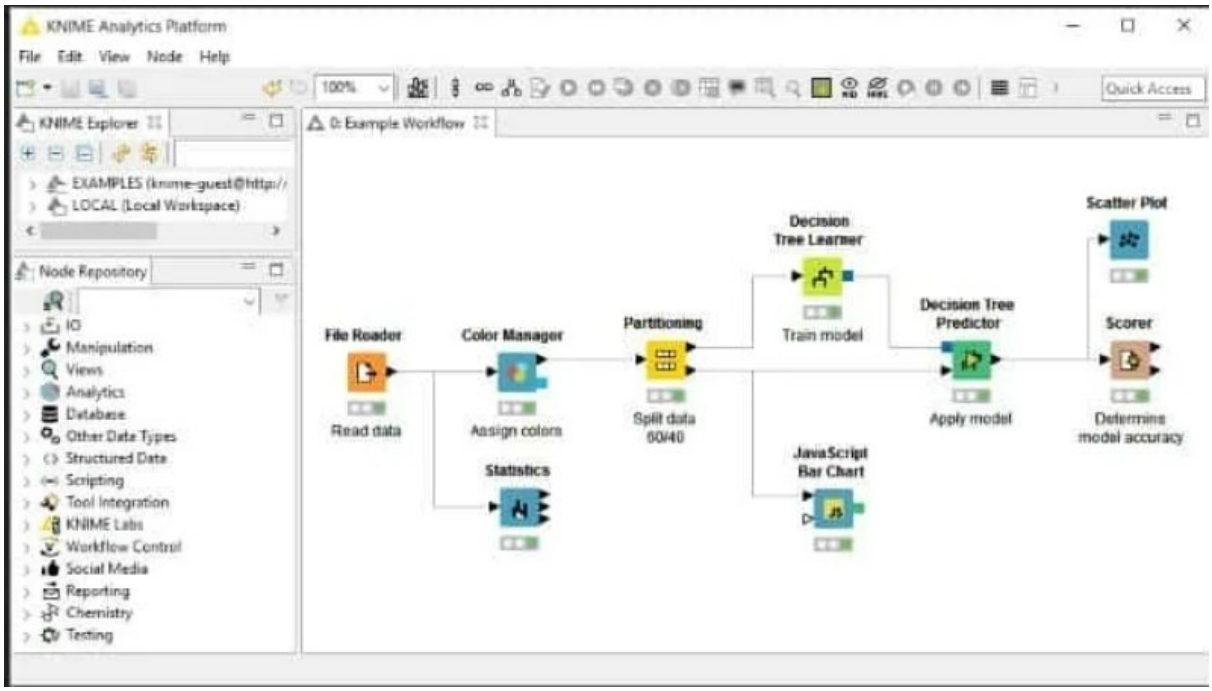

Рисунок 14 – Графічний інтерфейс Knime Analytics Platform

Має наступні особливості:

1. Потужна аналітика.
2. Синтез даних та інструментів.
3. Відкрита платформа.
4. Понад 1000 модулів, які постійно оновлюються.
5. Підтримка всіх основних баз даних та форматів файлів.
6. Підтримка типів даних: XML, JSON, зображення, документа та ін.
7. Математичні та статистичні функції.
8. Розширені алгоритми прогнозування та машинного навчання.
9. Керування робочим процесом.

10. Інструменти для Python, R, SQL, Java, Weka та ін.
11. Інтерактивні звіти.

До переваг системи Knime Analytics Platform можна віднести:

- аналіз відтоку;
- аналіз настроїв в соціальних мережах;
- кредитний скоринг.

Таблиця 9 – Рейтинг Knime Analytics Platform за обраними критеріями

Критерії оцінювання	Оцінка редактора	Оцінка користувачів (загальна)
Простота використання	8,6	6,3
Характеристики та функціональність	8,6	6,7
Розширені функції	8,5	8,0
Інтеграція з іншими системами	8,5	7,1
Продуктивність	8,4	7,0
Підтримка клієнтів	8,4	6,9
Простота реалізації		8,0
Оновлення та рекомендації		6,6
Загальний рейтинг	8,5	7,4

3.3. Платформа аналітики RapidMiner

Платформа RapidMiner – інтегроване середовище для машинного навчання, аналізу даних, аналізу текстів, прогнозування та бізнес-аналітики, яке використовується для бізнес-додатків, промислових систем, досліджень, освіти, навчання, швидкого створення прототипів та розробки додатків (рис. 15) [3]. Розповсюджується з комерційною ліцензією.

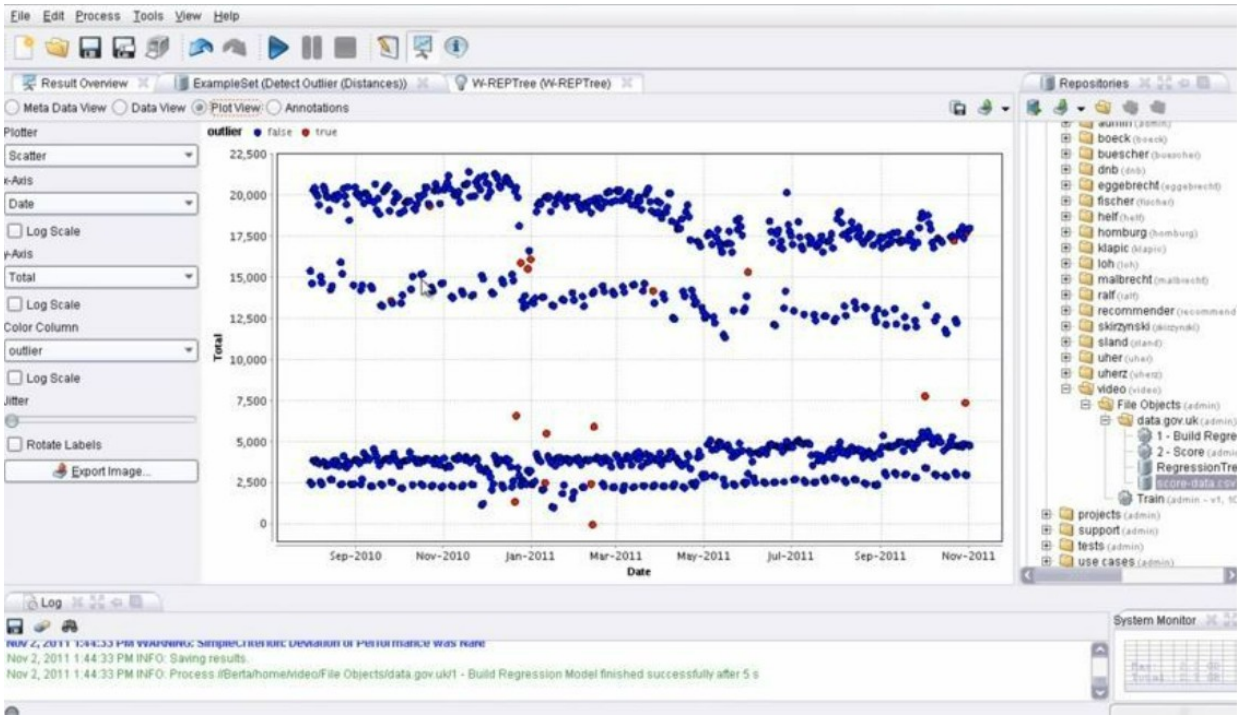

Рисунок 15 – Графічний інтерфейс RapidMiner

RapidMiner – це середовище візуального дизайну для швидкого створення повних прогнозних аналітичних робочих процесів. Має потужну бібліотеку алгоритмів машинного навчання, функцій підготовки та дослідження даних, а також інструментів перевірки моделей для підтримки проектів.

Має наступні особливості:

1. Уніфікована платформа.
2. Візуальний дизайн робочого процесу.
3. Широкі функціональні можливості.
4. Інновації з відкритим вихідним кодом.
5. Масштабованість даних.

До переваг системи RapidMiner можна віднести:

- детальна візуалізація робочого процесу;
- доповнення та оновлення з відкритим кодом;

- інтеграція з різними системами та підтримка різноманітних форматів.

Таблиця 10 – Рейтинг RapidMiner за обраними критеріями

Критерії оцінювання	Оцінка редактора	Оцінка користувачів (загальна)
Простота використання	9,4	6,2
Характеристики та функціональність	9,6	7,0
Розширені функції	9,4	8,0
Інтеграція з іншими системами	9,5	8,4
Продуктивність	9,5	7,9
Підтримка клієнтів	9,6	7,0
Простота реалізації		6,7
Оновлення та рекомендації		6,9
Загальний рейтинг	9,5	7,2

3.4. Результати оцінки програмних інструментів

В результаті проведеного аналізу обраних платформ аналізу даних за обраними критеріями отримали загальний рейтинг для кожної системи, які представлені в табл. 11. Отримані данні свідчать, що оцінка редактора значно вища ніж оцінка користувачів. Це обумовлено рівнем володіння навичками роботи з подібними системами. Але при виборі системи для використання в процесі подальших досліджень слід враховувати оцінку користувачів, оскільки вона більш об’єктивна тому, що складена за результатами опитування не однієї людина, а групи людей.

Таблиця 11 – Результати експертного оцінювання платформ

Критерії оцінювання	Оцінка редактора	Оцінка користувачів (загальна)
Weka	9,1	6,6
Knime Analytics Platform	8,5	7,4
RapidMiner	9,5	7,2

Оцінка редактора носить суб’єктивний характер і має бути розглянута тільки у якості інформаційної, тому для висновків її не слід брати до уваги.

Враховуючи вище зазначене та спираючись на дані наведені в табл. 11, можна зробити висновок, що із розглянутих систем найбільш ефективною є платформа Knime Analytics Platform.

4. ПРАКТИЧНЕ ВИКОРИСТАННЯ ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ ОБРОБКИ ДАНИХ ДЛЯ ПРОГНОЗУВАННЯ РОЗВИТКУ ПАНДЕМІЇЇ

4.1. Формування набору даних для аналізу

Розглянемо практичне застосування інструментальних засобів обробки даних на прикладі пандемії, яку викликав вірус COVID-19. Епідеміологічні дані про COVID-19 збираються та оприлюднюються Центром системних наук та інженерії університету Джона Хопкінса (JHU CCSE) [7]. Дані представлені в трьох окремих наборах даних для підтвердження зараження, одужання та випадки смерті, статистика оновлюються щодня, починаючи з 22 січня 2020 р. У кожному з цих наборів даних є запис (рядок) для кожного географічного регіону. Змінними у кожному наборі даних є провінція/штат, країна/регіон, широта, довгота та дата, починаючи з 22.01.20 р. Для кожного регіону значення для будь-якої дати вказує на сукупну кількість підтверджених випадків, випадків одужання та випадків смертей.

За методикою, що представлена в дослідженні [8], відповідно до вхідних вимог моделі змінили представлення даних таким чином, що замість трьох окремих наборів даних для трьох груп підтверджених, випадків одужання і смертельних випадків був організований лише один набір даних, що містить інформацію всіх трьох груп. В новому наборі даних, кожен запис (або рядок) набору даних містить інформацію про кількість підтверджених, випадків одужання або смертей за добу для кожного географічного регіону. В результаті, змінні в цьому новому наборі даних наступні: провінція/штат, країна/регіон, широта, довгота, дата (вказується певна дата, для якої представлені випадки). Кількість випадків із зазначенням кількості підтверджених, випадків одужання чи смертельних випадків на певну дату та визначення типу випадків представлені у вигляді, який запропоновано в дослідженні [9].

Для проведення аналізу були використані дані за 29 березня 2020 року з 50 660 записами та 7 змінними. Цей період включає інформацію про частину зими і весни в північній півкулі та частину літа і осіні у південній півкулі. До 29 березня набір даних складався із випадків із 177 країн та 252 різних регіонів світу. У наборі даних було 720 139 підтверджених, 33 925 смертей і 149 082 відновлених випадків.

4.2. Попередня обробка набору даних для аналізу

В дослідженні [10] запропонована методика попередньої обробки даних. Відповідно до неї попередня обробка даних здійснювалася над набором даних до етапу навчання представленої моделі. На рис. 16 показані етапи попередньої обробки. Спочатку набір даних було перевірено на наявність шуму, оскільки дані про шум вважалися негативними значення у змінній Cases. Набір даних містив 42 негативних значення в цій змінній. Після видалення цих значень, кількість записів було зменшено до 50 618.

В наборі даних, що розглядається, змінна Date була записана в числовому форматі та перейменована у змінну Day. У зв’язку з цим дата 22 січня 2020 р. вважається початком спалаху і наступні дні розраховувалися в розрізі днів від цієї дати, яка є початком координат. У зв’язку з цим дати 22 січня 29 березня 2020 р. вважалися днем 1 і днем 68, відповідно.

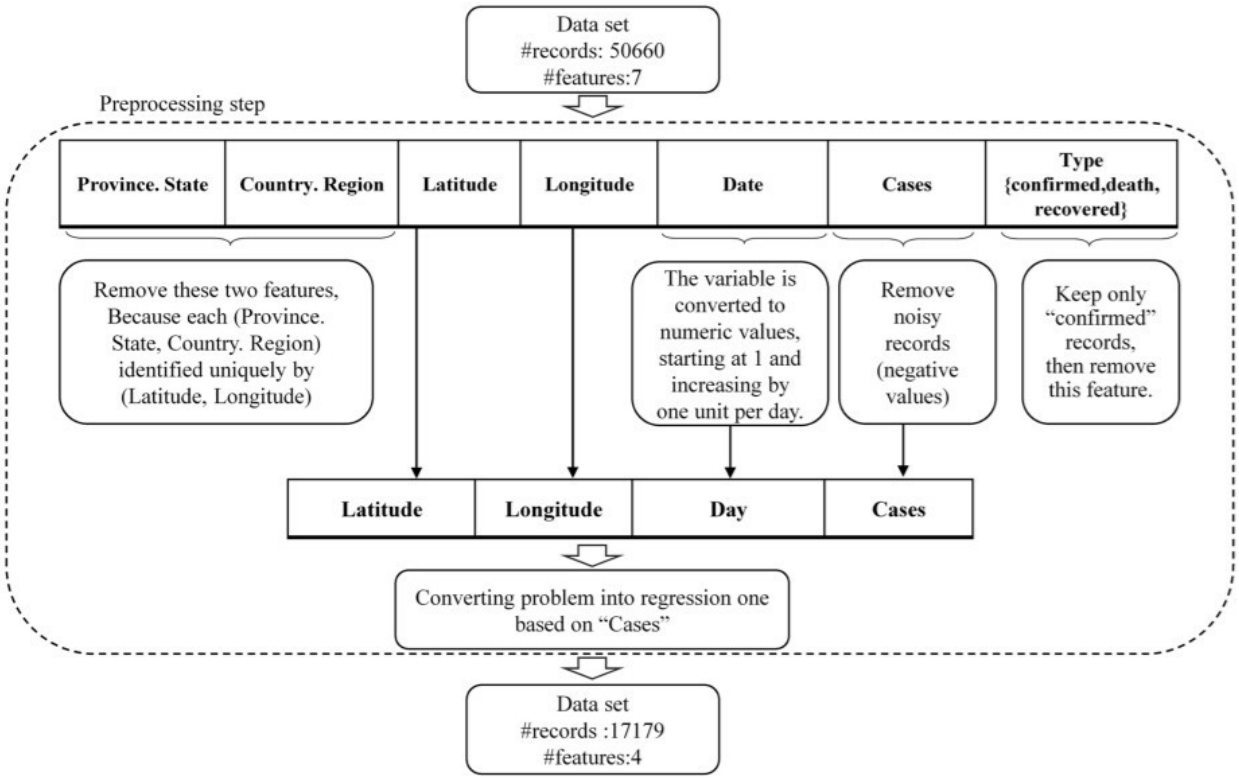

Рисунок 16 – Структурна схема попередньої обробки даних

Оскільки кожен регіон однозначно ідентифікується за своєю широтою та довгота, дані для провінції/штату та країни/регіон було виключено з набору даних. Крім цього, з урахуванням того, що робота спрямована на прогнозування саме захворюваності в будь-якому географічному регіоні, було розглянуто лише ті записи, що містять інформацію про підтверджені випадки захворювання (17 179 записів), без урахування випадків одужання та випадків смерті. Записи зі значенням «Підтверджено» у змінній Type, інші записи було видалено з набору даних, «Випадки» розглядалися як залежна змінна.

4.3. Модель прогнозування

Для побудови моделі прогнозування використовується ансамблевий метод регресії навчання, який дозволяє спрогнозувати захворюваність на COVID-19 у різних регіонах. Ідея ансамблевого навчання полягає в побудові

моделі шляхом поєднання сильних сторін набору простіших базових моделей, які називаються слабкими учнями [11]. На кожному кроці, ансамбль підходить новому учню до різниці між реакцію та агрегованим прогнозом усіх учні, які навчалися раніше. Один із найпоширеніших методів визначення функції втрат є визначення помилки методом найменших квадратів (LS) [12].

В запропонованій моделі використовувався набір учнів, які намагаються мінімізувати середню квадратичну помилку (MSE) методом найменших квадратів (LSBoost). Результат моделі Fm(x) на кроці m розраховується за допомогою рівняння (1):

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m) \tag{1}$$

де x – вхідна змінна, а h (x;a) – функція x, що характеризується параметрами моделі.

Значення ρ розраховуються за допомогою наступного рівняння (2):

$$(\rho_m, a_m) = \arg min_{a,\rho} \sum_{i=1}^{N} [\tilde{y}_i - \rho h(x_i; a)]^2 \tag{2}$$

де N – кількість навчальних даних, a_i – різниця між реакцією та агрегованим прогнозом на попередньому кроці.

Структурна схема моделі прогнозування представлена на рис. 17.

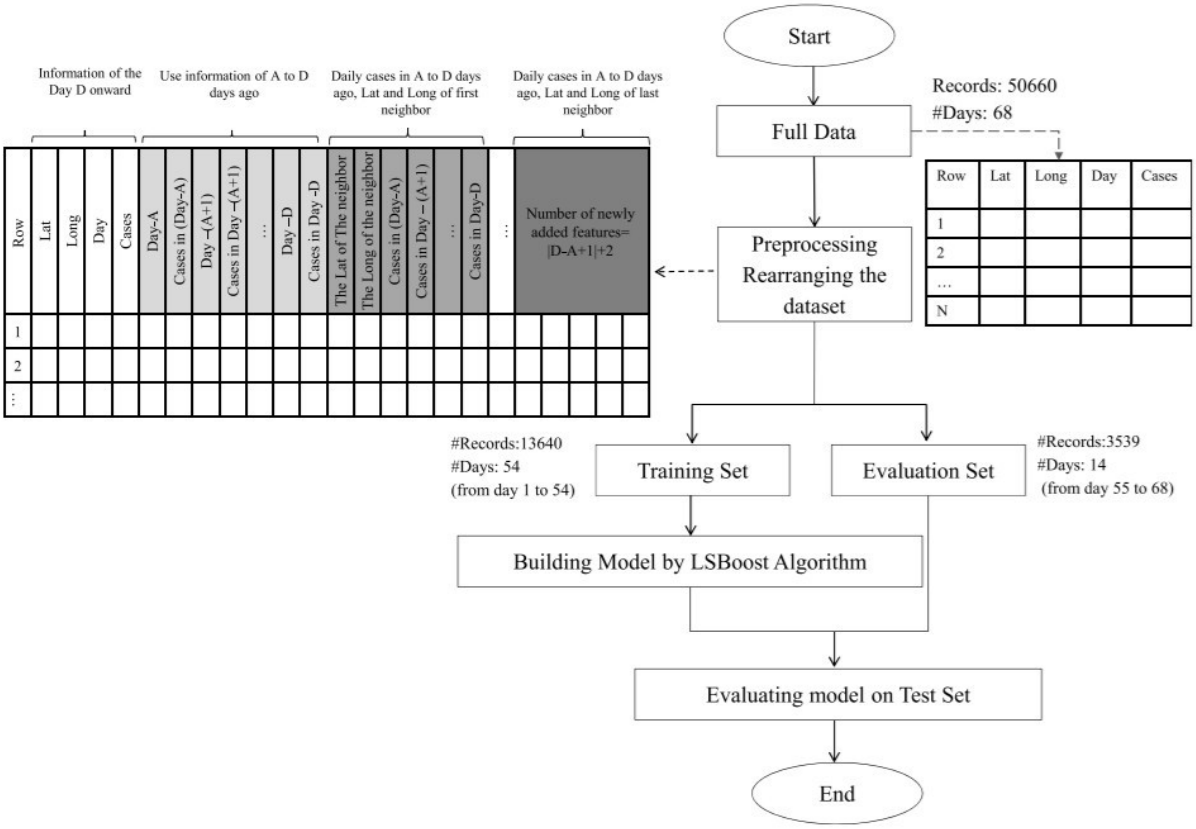

Рисунок 17 – Структурна схема моделі прогнозування

Так як інкубаційний період COVID-19 приблизно становити 14 днів, в роботі було зроблено припущення, що потрібно щонайменше 14 днів попередньої інформації для прогнозування захворюваності на Covid-19 за 1 день. Тому запропонована модель розглянула все можливі інтервали між першими і останніми 14 днями, щоб знайти оптимальний період часу для використання цієї інформації для прогнозування кількості випадків у найближчі дні. Було зроблено припущення, що захворюваність у будь-якому регіоні може відбуватися за схемою останніх днів у тому самому регіоні та поблизу. Тому після визначення оптимального періоду часу, модель додала інформацію про підтверджені випадки в кожному регіоні та поблизу в зазначеному періоді до того самого запису регіону в набір даних.

Після встановлення інтервалу часу [A, B] та числа сусідів, набір даних був переупорядкований. Кількість записів була зменшена з N до M, де M розраховується за формулою (3):

$$M = N - (B \times R) \quad (3)$$

де R – кількість різних регіонів у наборі даних, а B – останній день періоду часу.

При цьому збільшилася кількість змінних, що зберігаються для кожного запису з перших 4 змінних (широта, довгота, день і випадки) до F, яке розраховується за формулою (4):

$$F = 4 + 2 \times |B-A+1| + NN \times (|B-A+1| + 2) \quad (4)$$

де NN – кількість сусідів, а 4 – кількість змінних у вихідному наборі даних, оскільки для кожного географічного регіону зберігаються широта, довга, день та випадки. |B-A + 1| – кількість днів у періоді, які беруть участь у прогнозі наступних 14 днів.

Значення NN множиться на 2 оскільки для кожного сусіда до інформації запису додаються широта та довгота. Крім того, для кожного дня в межах періоду прогнозування кількість випадків була додана до інформації запису, тому NN було помножено на |B-A + 1|. Для кожного регіону були додані дані про день і випадки за період. Таким чином, |B-A + 1| було помножено на 2. Однак слід зазначити, що залежною змінною залишилися «Випадки поточного дня».

Оскільки кількість найближчих регіонів і попередніх днів, ефективних для прогнозування, були невідомі, було прийнято, що ці значення невідомі змінні. Таким чином отримали найточнішу модель, дослідивши всі можливі комбінації таких змінних в ітераційному процесі.

Точність моделі була оцінена методами середньоквадратичної помилки (MSE) і середньої абсолютної похибки (MAE). Завдяки нормалізації MAE між [0,1] похибка оцінки дорівнює 2-кратному MAE. Для цього інформацію за останні 2 тижні по всіх регіонах розглядали як набір для перевірки, а модель навчали з використанням іншої інформації в наборі даних.

Прогнозування захворюваності для найближчих двох тижнів здійснювали наступним чином. Створили новий набір тестів для прогнозування захворюваності протягом наступних 2 тижнів (до 12 квітня 2020 року). Кількість записів у цьому наборі даних дорівнювала кількості унікальних географічних регіонів у наборі даних COVID-19. Потім, відповідно до найкращого сусідства та оптимального інтервалу часу, вказаних на попередньому кроці, для кожного запису були надані необхідні характеристики. Після цього найкращу модель, створену на попередньому кроці, перенавчали на всьому наборі даних як навчальний набір. Пізніше ця модель була досліджена на новому наборі тестів для прогнозування рівня захворюваності.

На наступному етапі виконали оцінку фактичної роботи запропонованої моделі. Враховуючи, що фактична кількість підтверджених випадків за період з 30 березня по 12 квітня 2020 року була доступна на момент огляду, ефективність запропонованої моделі була визначена на основі відсоткової похибки між прогнозованими та фактичними значеннями. Відсоткова помилка була розрахована за формулою (5):

$$\delta = \left(\frac{|v_A - v_E|}{v_A}\right) \times 100 \quad (5)$$

де δ – відсоткова помилка, v_A – це фактичне спостережуване значення, а v_E – очікуване (передбачене) значення.

Крім того, відповідно до прогнозованих та фактичних підтверджених випадків у 252 географічних регіонах у наборі даних, рівень континентальної захворюваності був розрахований за допомогою рівняння (6):

$$Continental\ incidence\ rate = \left(\frac{I_C}{I_W}\right) \times 100 \quad (6)$$

де I_C – захворюваність на кожному континенті, а I_W – це глобальна захворюваність на COVID-19 з 30 березня по 12 квітня 2020 року.

4.4. Результати використання моделі прогнозування

Кількість сусідів коливалася від нуля до 10. Значення 10 було отримано методом підбору. Для обчислення найближчих сусідів використовувалася евклідова відстань на основі широти та довготи. Враховуючи, що набір даних містить дані з 22 січня 2020 року по 29 березня 2020 року за день, коли необхідно визначити прогнозоване значення захворюваності, найближчий і найдальший днів було обрано 14 і 54 відповідно. Оскільки кількість підтверджених випадків досить різна від регіону до регіону, запропонований алгоритм було впроваджено для 3 різних груп регіонів: для регіонів з менш ніж 200 підтвердженими випадками на добу (16 825 записів), з 200 до 1000 випадків на добу (220 записів), а також ті, в яких реєструється понад 1000 випадків на день (152 записи).

В таблиці 12 наведено результати найкращої запропонованої моделі з урахуванням різного складу мікрорайону та днів. Для прогнозування захворюваності на COVID-19 у регіонах з понад 1000 підтверджених випадків на добу запропонована модель продемонструвала найкращі показники з MAE 6,13%, враховуючи інформацію за останні 14-17 днів регіону та два його сусідні райони. У наборі даних кількість зареєстрованих випадків у цих регіонах варіювалася від 1019 до 19 821.

Таблиця 12 – Результати моделювання захворюваності

Maximum number of confirmed cases in a day		Number of Neighbors	Interval of days [min, max]	MSE		MAE	
				Value	Percent	Value	Percent
< 200	Train	–	[14,34]	1.86	0.005%	0.52	0.29%
	Test			407.47	1.04%	9.12	4.71%
[200,1000)	Train	9	[14, 20]	1.71	0.002%	0.62	0.07%
	Test			1.59e+ 04	1.87%	79.01	8.54%
≥1000	Train	2	[14, 17]	140.62	0.00003%	5.89	0.03%
	Test			7.14e+ 06	1.79%	1.2e+ 03	6.13%

Для регіонів із 200 – 1000 випадками на добу запропонована модель показала найкращі результати щодо 9 найближчих сусідніх областей та з даними за останні 14–20 днів із MAE 8,54% на наборі перевірки. З іншого боку, для регіонів з менш ніж 200 випадками на добу запропонована модель працює найкраще з MAE 4,71%, враховуючи дані регіону за останні 14-34 дні.

Результати прогнозування захворюваності до 12 квітня 2020 року отримали також з використанням цієї моделі. Отримали значення, що характеризують поширеність COVID-19 з першого по десятий тиждень у різних регіонах на основі інформації, наданої епідеміологічним набором даних COVID-19 [7]. Графічно діаметр кіл пропорційний поширенню в цих регіонів, а центр кожного кола відповідає географічним координатам регіону.

В таблиці 13 наведено результати прогнозу щодо кількості нових випадків за добу на різних континентах. Відповідно до розташування материків у північній та південній півкулях, період, про який йде мова, містить інформацію про зиму та ранню весну на континентах Північної Америки, Європи та майже в усіх частинах Азії. Він включає літо та частину осені в Австралії та приблизно всієї Південної Америки.

Враховуючи, що Африка розташована у всіх чотирьох півкулях, дані, записані для цього континенту в цей період у наборі даних, включають усі пори року.

Таблиця 13 – Результати моделювання захворюваності для континентів

Date	*Continents*						*Total number of confirmed cases*
	Africa	**Asia**	**Australian**	**Europe**	**North America**	**South America**	
22 Jan ~ 29 Mar	4995	161,986	4522	385,097	150,877	11,740	719,217
30-Mar	635	7720	802	37,853	19,269	1906	68,185
31-Mar	820	7227	722	37,433	16,890	2000	65,092
1-Apr	472	7533	338	38,512	19,625	1508	67,988
2-Apr	1046	6438	981	44,047	18,435	1955	72,902
3-Apr	1047	6790	780	53,087	19,802	2359	83,865
4-Apr	1015	9739	872	51,954	19,302	2258	85,140
5-Apr	1014	10,563	1226	47,352	19,579	2490	82,224
6-Apr	1447	6867	1015	48,562	19,060	2530	79,481
7-Apr	1636	8027	1057	51,192	20,191	2768	84,871
8-Apr	2087	6786	1444	56,826	19,546	2550	89,239
9-Apr	2157	7749	1270	55,316	20,475	2685	89,652
10-Apr	1976	5818	1430	54,377	20,819	2573	86,993
11-Apr	1849	8962	1390	56,284	19,627	2351	90,463
12-Apr	1930	6781	1199	54,870	20,337	2806	87,923
Total	19,131	107,000	14,526	687,665	272,957	32,739	1,134,018
Prevalence growth rate	**283.00**	−33.94	221.23	78.57	80.91	178.87	57.67

Очікувалося, що до 12 квітня в усьому світі буде зареєстровано 1 134 018 нових випадків. З них найпоширенішими були Європа з 687 665 (60,64%), Північна Америка з 272 957 (24,07%) та Азія з 107 000 (9,44%) випадками, тоді як Австралія з 14 526 (1,28%), Африка з 19 131 (1,69%) та Південна Америка з 32 739 (2,89%) новими випадками були найменшими. Найвищі показники захворюваності на COVID-19 були в Африці, Європі та Південній Америці. Азія була єдиним континентом, який уповільнив своє зростання з показником захворюваності − 34. На рис. 18, 19 показано прогноз рівня захворюваності в різних регіонах. Відповідно прогнозу поширеність повинна була зменшитися протягом наступних 2 тижнів на Близькому Сході, але збільшитися в Північній Америці та Європі.

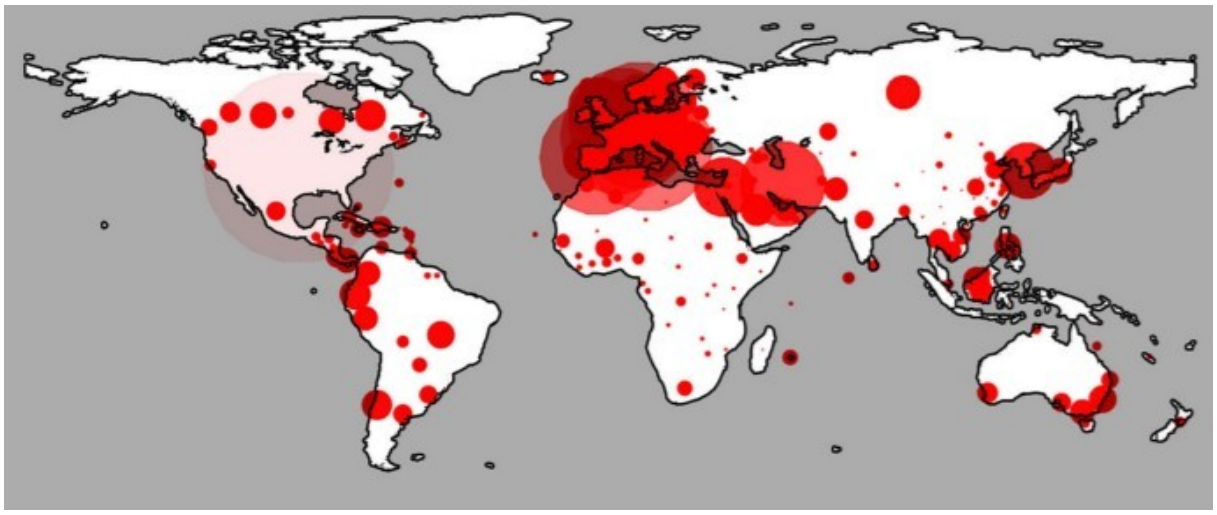

Рисунок 18 – Прогноз рівня захворюваності на 75 день

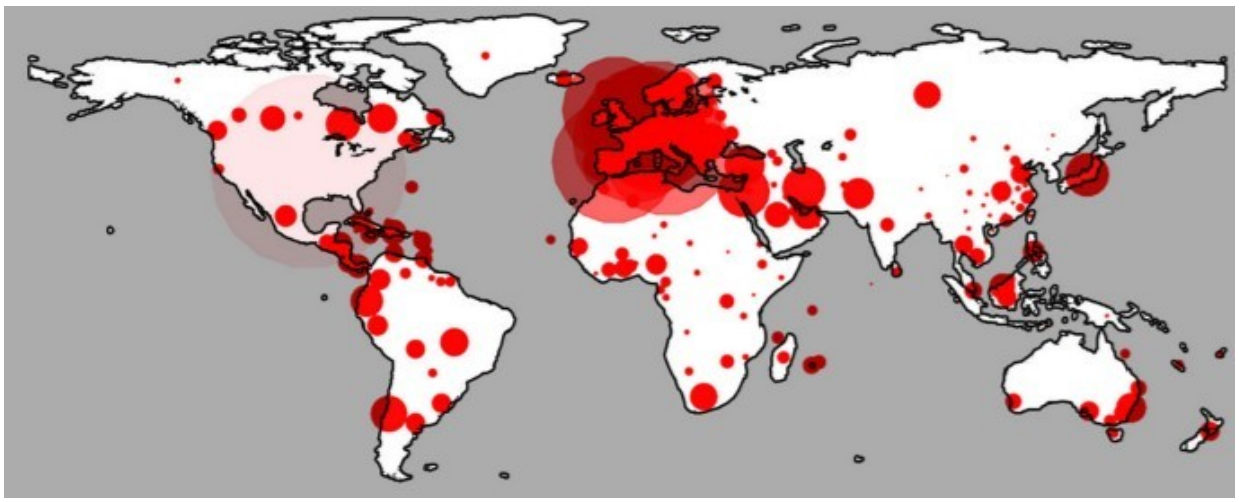

Рисунок 19 – Прогноз рівня захворюваності на 82 день

На наступному етапі дослідження було проведено порівняння прогнозованих і фактичних випадків від 30 березня до 12 квітня 2020 року. В таблиці 3 наведено загальну кількість щоденних випадків у 252 регіонах, які досліджувалися з 30 березня по 12 квітня 2020 року.

Таблиця 14 – Результати порівняння прогнозованої захворюваності та фактичної

Date	Across all 252 geographic regions		Percent error
	Predicted	Actual	
30-Mar	68,185	65,321	4.38%
31-Mar	65,092	76,799	15.24%
1-Apr	67,988	76,657	11.31%
2-Apr	72,902	81,340	10.37%
3-Apr	83,865	83,272	0.71%
4-Apr	85,140	80,392	5.91%
5-Apr	82,224	71,994	14.21%
6-Apr	79,481	73,285	8.45%
7-Apr	84,871	77,773	9.13%
8-Apr	89,239	84,275	5.89%
9-Apr	89,652	86,461	3.69%
10-Apr	86,993	87,520	0.60%
11-Apr	90,463	76,217	18.69%
12-Apr	87,923	95,353	7.79%
Total number of confirmed cases	1,134,018	1,116,659	1.55%

Дані, що наведені в таблиці, показують, що щоденна відсоткова похибка становить менше 20%. Найкраща точність запропонованої моделі в прогнозуванні захворюваності на COVID-19 була отримана 10 квітня (99,6%), а найгірша 11 квітня (81,3%). Аналіз даних двотижневих континентальних показників захворюваності показано на рис. 20, 21.

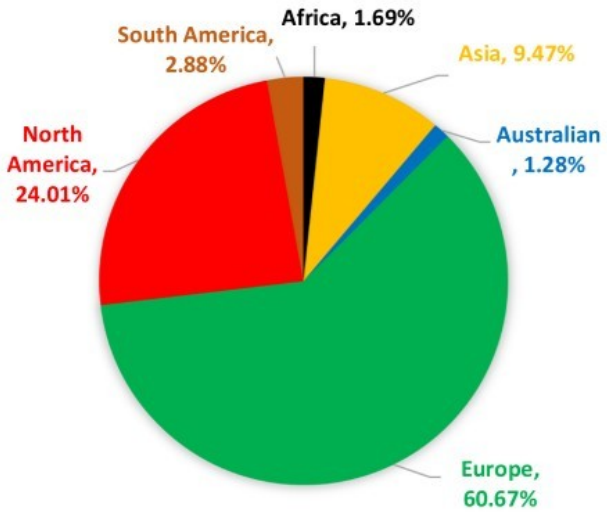

Рисунок 20 – Прогнозований рівень захворюваності

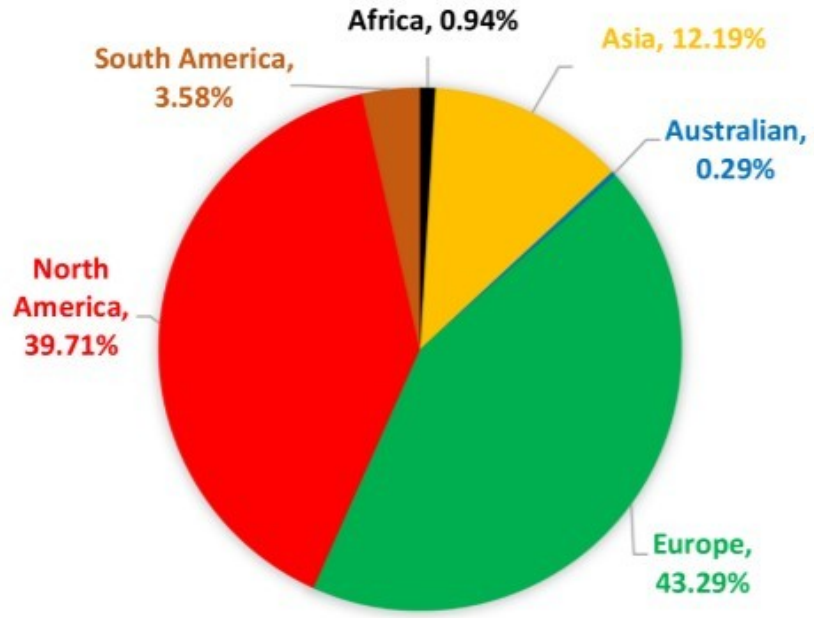

Рисунок 21 – Фактичний рівень захворюваності

Найкращі прогнозовані показники континентальної захворюваності були виявлені в Південній Америці та Азії з похибкою 18,15 та 21,04% відповідно. Найгірші випадки, як і раніше, спостерігалися в Африці та Австралії з більш ніж 80% помилок.

Проведені дослідження [10] показали, що інтелектуальний аналіз даних здатний представити прогнозну модель і витягти нові знання з ретроспективних даних. Спосіб обробки даних, а також обрані змінні мали значний вплив на відкриття знань. Для прогнозування спалаху використовуються різні методи аналізу даних. Як фактична глобальна проблема охорони здоров'я, COVID-19 вже перетворився на одну з найбільших надзвичайних ситуацій у світі. В дослідженні було запропоновано дослідити його спалах у всьому світі протягом двотижневого періоду за допомогою прогнозної моделі, заснованої на ретроспективних даних. Отримали висновок, що така модель може бути представлена з прийнятними показниками помилок. У дослідженні використовувався набір даних про коронавірус для розробки моделі прогнозування захворюваності

на COVID-19. Відповідно до коефіцієнта захворюваності на добу, модель була підготовлена на основі трьох груп – менше 200, 200 – 1000 та понад 1000 випадків. Результати одностороннього дисперсійного аналізу показали, що існує статистично значуща різниця між показниками поширеності у трьох групах (p-значення < 0,001). Для кожної групи була реалізована модель прогнозування та прогнозована захворюваність на наступні 2 тижні. Запропонована модель досягла похибки близько 10% (90% точності) для групи менше 200 випадків, 18% похибки (82% точності) для групи 200 – 1000 випадків і 13% похибки (87% точності) для групи понад 1000 випадків.

Оскільки захворюваність на COVID-19 оцінювалася протягом 68 днів у всьому світі, а модель прогнозу була представлена на двотижневий період (тобто 30 березня – 12 квітня 2020 р.), очікувалося, що понад 1 000 000 людей можуть захворіти протягом наступних 2 тижнів, що за статистикою зросло на 58% порівняно з 700 000 випадків спалаху станом на 29 березня 2020 року.

Дослідження виявило, що сусідні регіони з поширеністю менше 1000 мали подібну захворюваність, тому захворюваність кожного з цих регіонів можна було визначити на основі інформації про сусідні області. Використання інформації про граничні регіони дає змогу моделі опосередковано враховувати ефективну політику інших регіонів у прогнозуванні захворюваності на COVID-19 у кожному регіоні.

Враховуючи, що запропонована модель була навчена з використанням лише 68-денних даних (які були найактуальнішою інформацією на момент написання), точність прогнозування захворюваності вище 81% вважалася прийнятною для такого невідомого захворювання. Далі, згідно з результатами, наведеними в табл. 3, модель похибки прогнозу для сумарного з 12 днів для 252 регіонів становила менше 2%. Хоча від нової пандемії очікується багато невідомих нових умов, ця інформація може керувати плануванням та розподілом ресурсів для профілактики, лікування та медичної допомоги.

Хоча часові ряди зазвичай мають бути достатньо довгими (зазвичай кілька років), щоб адекватно врахувати сезонність, на основі результатів реалізації моделі, можна сказати, що ця модель, навіть з таким коротким часовим рядом, здатна керувати сезонністю та може передбачити кількість випадків із прийнятною точністю. Проте пропонується, щоб майбутні дослідження спеціально розглядали вплив сезонних змін на поширеність цього захворювання.

Одним з обмежень дослідження було те, що набір даних не надав достатньої інформації з усіх континентів. Враховуючи, що захворювання виникло не одночасно на всіх континентах, а континентальна поширеність була в більшості випадків після 40-го дня після першого випадку в Китаї дані за 68 днів не здавалися достатніми, щоб передбачити поширеність такого невідомого захворювання.

В Африці перший випадок був зареєстрований у понад 80% із 45 географічних регіонів з 50-го дня. Кількість підтверджених випадків з того часу склала 4682, що становило 97,83% від загальної кількості 4783 підтверджених випадків в Африці. В Австралії перший випадок був зареєстрований у понад 45% з 11 географічних регіонів, починаючи з 40-го дня. Також із загальної кількості 4504 випадків на континенті тоді було підтверджено 4478 випадків (99,4%).

В Європі перший випадок був зареєстрований у 60 із 69 географічних регіонів у наборі даних, починаючи з 40-го дня. Із 385 735 випадків з цього дня також введено інформацію про 384 268 випадків (тобто 99,62%). Так само, Південна Америка підтвердила свій перший випадок після 40-го дня в 16 з 17 регіонів. Очевидно, що із загальної кількості 11 642 випадків 11 542 (14,99%) були підтверджені з 50-го дня. Навпаки, у 88% північноамериканських регіонів були підтверджені перші випадки з 50 дня. Крім того, з 46 підтверджених випадків на континенті станом на 29 березня 2020 року, 38 були зареєстровані з 50 дня (82,61%) і 41 був підтверджений. з 40 дня (89,13%). Через недостатню інформацію про деякі континенти через їх

поширення пізніше заявленого початку спалаху, дія таких заходів, як збільшення кількості тестів, що проводяться щодня, а також карантинних обмежень на деяких континентах, таких як Європа, починається в місця з 30 березня по 12 квітня, не були відображені в наборі даних.

Тим не менш, неточне прогнозування кількості випадків в Африці, у свою чергу, можна пояснити недостатньою інформацією про континент у наборі даних. У 80% африканських регіонів перший підтверджений випадок був зафіксований через 50 днів після спалаху. Із 4786 випадків до 68-го дня зареєстровано 4682 випадки (більше 97%) з 50-го дня.

Крім того, через те, що широта і довгота є двома важливими показниками в наборі даних, нерівномірність запису цієї інформації для різних географічних регіонів є ще одним обмеженням роботи; для деяких областей інформація стосується одного штату країни, а для деяких областей – для всієї країни.

Наприклад, у наборі даних для США всі випадки надаються лише з точки зору однієї широти та довготи, але для Нідерландів дані про випадки COVID-19 надаються для чотирьох різних пар широти та довготи. Іншим обмеженням проведеного дослідження було використання даних з всі країни, які борються з COVID-19, мають власні протоколи тестування та ідентифікації пацієнтів. Однак загалом це єдиний глобальний набір даних для COVID-19, який використовувався в інших дослідженнях [13, 14]. Крім того, у запропонованій моделі було враховано попередню інформацію про кожну країну для прогнозування захворюваності в цій країні, щоб зменшити зазначене обмеження.

Варто зазначити, що модель спирається як на інформацію, надану набором даних, так і на поточні заходи, вжиті для боротьби з хворобою. Отже, якщо політика уряду щодо боротьби з хворобою зміниться, зміниться і точність інформації.

Проведене дослідження спиралося на дані з 22 січня по 29 березня, надані Університетом Джона Хопкінса, і запропонувало більш складну

модель, засновану на машинному методі навчання. Середня абсолютна похибка запропонованої моделі склала 6,13% при прогнозуванні захворюваності на COVID-19 за двотижневий період 16 – 29 березня для регіонів із понад 1000 випадків за добу. MAE становив 8,45 та 4,71% для регіонів із добовою захворюваністю від 200 до 1000 випадків та менше 200 випадків відповідно. Точність більш ніж 82% оцінки підтверджує уявлення про те, що на структуру захворюваності в регіоні впливає картина захворювання за останні дні в тому ж регіоні та сусідніх областях.

Незважаючи на численні обмеження набору даних, відсутність знань про таку невідому хворобу та зміни в політиці боротьби з хворобами в різних країнах протягом досліджуваного періоду, запропонована модель виявилася ефективною для прогнозування глобальної захворюваності на COVID-19 у двотижневий період 30 березня та 12 квітня з точністю 98,45%. Крім того, точність запропонованої моделі в прогнозуванні щоденних випадків в найгіршому сценарії становила 81,31%.

ВИСНОВКИ

В роботі було розглянуто інструментальні засоби інтелектуального аналізу даних та проведені дослідження функціональних можливостей таких платформ. За результатами проведеного дослідження були визначені найбільш ефективних для використання в соціально-екологічних інформаційних системах.

На першому етапі було розглянуто та коротко описано 11 програмних інструментів аналізу даних. За результатами проведеного аналізу біло обрано п’ять платформ, а саме, Python, R, Weka, Knime, RapidMiner, як найбільш задовольняючі висунутим умовам. Далі були визначені критерії для порівняльного аналізу обраних систем. В результаті проведеного аналізу отримали загальну оцінку для кожної системи, яка визначалась як сума оцінок по критеріям.

В результаті проведеного порівняльного аналізу платформ для аналізу даних не вдалося визначити найбільш ефективну систему для вирішення складних соціально-екологічних задач. Тому було прийнято рішення провести додаткове дослідження платформ Weka, Knime, RapidMine з метою виявлення найбільш ефективної системи. Методом дослідження було обрано експертну оцінку тому, що простий порівняльний аналіз не дав бажаних результатів.

На наступному етапі отримали загальний рейтинг для кожної системи. Отримані данні свідчать, що оцінка редактора значно вища ніж оцінка користувачів. Це обумовлено рівнем володіння навичками роботи з подібними системами. Але при виборі системи для використання в процесі подальших досліджень слід враховувати оцінку користувачів, оскільки вона більш об’єктивна тому, що складена за результатами опитування не однієї людина, а групи людей.Враховуючи вище зазначене та спираючись на

отримані дані, можна зробити висновок, що із розглянутих систем найбільш ефективною є платформа Knime Analytics Platform.

Заключним етапом роботи було представлення результатів практичного використання систем інтелектуального аналізу даних на прикладі побудови моделі для прогнозування поширення епідемій.

ПЕРЕЛІК ПОСИЛАНЬ

1. Барсегян, А. А. Анализ данных и процессов: учеб. пособие. – СПб.: БХВ-Петербург, 2009. – 512 с.

2. Halkidia M., Spinellisb D., Tsatsaroniscand G., Vazirgiannisc M. Data Mining in Software Engineering // Intelligent Data Analysis, 15 (2011). p. 413-441.

3. Data Mining Software / https://www.predictiveanalyticstoday.com/top-free-data-mining-software/ (дата звернення 28.11.2021).

4. 25 найкращих інструментів для інтелектуального аналізу даних / https://coderlessons.com/tutorials/bolshie-dannye-i-analitika/teoriia-khraneniia-dannykh/31-25-luchshikh-instrumentov-dlia-intellektualnogo-analiza-dannykh (дата звернення 28.11.2021).

5. Пять ключевых библиотек и пакетов для анализа данных на Python / https://techrocks.ru/2018/07/22/5-key-libraries-and-packets-for-data-analysis-in-python/ (дата звернення 28.11.2021).

6. Hawkins M. Extending KNIME Python Integration with Plotly Express and Kaleido / https://towardsdatascience.com/tagged/knime-analytics-platform (дата звернення 28.11.2021).

7. CCSE. J.H.U.C.f.S.S.a.E.J. Novel Coronavirus (COVID-19) Cases Data. 2020 / https://data.humdata.org/dataset/novel-coronavirus-2019- ncov-cases (дата звернення 02.12.2021).

8. Predicting the incidence of COVID-19 using data mining Fatemeh Ahouz1 and Amin Golabpou? Ahouz and Golabpour BMC Public Health (2021) 21:1087 https://doi.org/10.1186/s12889-021-11058-3.

9. Krispin R. Coronavirus. 2020. Available / https://github.com /RamiKrispin/ coronavirus (дата звернення 02.12.2021).

10. Predicting the incidence of COVID-19 using data mining Fatemeh Ahouz1 and Amin Golabpou? Ahouz and Golabpour BMC Public Health (2021) 21:1087 https://doi.org/10.1186/s12889-021-11058-3.

11. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, second edition. Springer Series in Statistics. New York: Springer-Verlag; 2008, p. 512.

12. Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat. 2000;29:1189 –232. https://doi.org/10.1214/aos/1013203451.

13. Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analyzing the epidemiological outbreak of COVID-19: a visual exploratory data analysis approach. J Med Virol. 92(6):632–8. https://doi.org/10.1002/jmv.25743.

14. Binti Hamzah FA, et al. CoronaTracker: world-wide COVID-19 outbreak data analysis and prediction. 2020.