КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ РАДІОФІЗИКИ, ЕЛЕКТРОНІКИ ТА КОМП'ЮТЕРНИХ СИСТЕМ

Кафедра радіотехніки та радіоелектронних систем

До захисту допущено: «На правах рукопису»
Завідувач кафедри _____ Ігор АНІСІМОВ
« __ » червня 2023 р.

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему:

«ІДЕНТИФІКАЦІЯ ХЛОРОРГАНІЧНИХ МОЛЕКУЛ МЕТОДОМ МАШИННОЇ КЛАСИФІКАЦІЇ ЇХ МАС-СПЕКТРІВ»

Виконав:
студентка 4-го курсу
денної форми навчання
спеціальності 172 - Телекомунікації та радіотехніка
ОП «Інформаційна безпека телекомунікаційних систем і мереж»
Шелякіна Богдана-Марія Анатоліївна _____

Науковий керівник:
д.т.н., доц. Ольшевський Сергій Валентинович _____

Рецензент:
к.х.н., с.н.с. Кофанов Валерій Іванович _____

Засвідчую, що у цій бакалаврській роботі
немає запозичень з праць інших авторів без
відповідних посилань
Студент _____

Робота допущена до захисту в ЕК рішенням кафедри радіотехніки та радіоелектронних систем від «22» червня 2023 р., протокол № 21.

Завідувач кафедри радіотехніки та радіоелектронних систем,
доктор фіз.-мат. наук, професор
Анісімов Ігор Олексійович _____

Київ – 2023

ЗМІСТ

ВСТУП

Ідентифікація хлорорганічних молекул є важливим завданням в хімічному аналізі і дослідженнях, оскільки ці сполуки можуть використовуватись у багатьох галузях, включаючи фармацевтику, пестициди, полімери та багато інших промислових процесів. Один з найефективніших способів ідентифікації хлорорганічних молекул - це застосування методу машинного навчання для класифікації їх мас-спектрів.

Мас-спектрометрія є потужним інструментом для аналізу хімічних речовин, зокрема хлорорганічних сполук. Мас-спектри надають інформацію про маси та заряди іонів, утворених в результаті розпаду молекули на фрагменти під дією енергетичного збудження. Кожна хлорорганічна молекула має унікальний мас-спектр, що дозволяє відрізнити її від інших сполук.

Методи машинного навчання, зокрема класифікація, забезпечують здатність автоматично визначати типи сполук на основі навчання на великій кількості попередньо класифікованих мас-спектрів. Ці алгоритми навчаються розпізнавати характерні ознаки і закономірності в мас-спектрах хлорорганічних сполук, що дозволяє їм з високою точністю класифікувати невідомі зразки.

Використання методу машинного навчання для ідентифікації хлорорганічних молекул має безліч переваг, таких як швидкість, точність та здатність обробляти великі обсяги даних. Цей підхід відкриває нові можливості для виявлення нових хлорорганічних сполук, вивчення їх хімічних властивостей та дослідження їх ролі в різних процесах.

Одним з основних викликів у використанні методу машинного навчання для класифікації мас-спектрів хлорорганічних молекул є створення відповідних навчальних наборів даних. Ці набори повинні містити достатню кількість представників кожного класу сполук для навчання алгоритмів та досягнення надійних результатів.

Додатковим викликом є врахування варіабельності мас-спектрів, яка може бути спричинена різними чинниками, такими як умови аналізу, джерело зразків і технічні характеристики мас-спектрометра. Розробка стійких до таких

варіацій моделей машинного навчання є важливим аспектом дослідження в цій області.

Незважаючи на ці виклики, метод машинної класифікації мас-спектрів є потужним інструментом для ідентифікації хлорорганічних молекул. Він сприяє прискоренню і автоматизації процесу аналізу, дозволяючи швидше і ефективніше виявляти, класифікувати та вивчати хлорорганічні сполуки у різних областях науки і промисловості.

РОЗДІЛ 1. ПОСТАНОВКА ЗАДАЧІ

Метою нашої теми є розробка і застосування методу машинної класифікації мас-спектрів для ідентифікації хлорорганічних молекул. Цей метод дозволяє автоматично аналізувати мас-спектри та класифікувати їх на основі характерних особливостей, що дозволяє ефективно і швидко визначати типи хлорорганічних сполук.

Актуальність роботи полягає в тому, що хлорорганічні сполуки є широко поширеними і мають важливе значення у багатьох галузях, включаючи фармацевтичну, агрохімічну, пластикову та хімічну промисловість. Проте, вони також можуть бути токсичними та небезпечними для навколишнього середовища, тому контроль їх присутності є важливим завданням для забезпечення безпеки та якості продуктів.

Традиційні методи ідентифікації хлорорганічних молекул, такі як мас-спектрометрія, вимагають експертного аналізу та великої кількості часу для обробки даних. Використання методів машинного навчання і класифікації дозволяє автоматизувати процес ідентифікації, зменшити час аналізу та знизити ймовірність помилки.

Розробка ефективного методу машинної класифікації мас-спектрів хлорорганічних молекул має значний потенціал у різних галузях науки та промисловості. Вона може допомогти в розробці нових лікарських препаратів, виявленні забруднень у воді та продуктах харчування, контролі якості хімічних речовин та в інших областях, де важлива точна та швидка ідентифікація хлорорганічних сполук.

Спочатку ми зосереджуємося на проведенні аналізу інформації з доступних джерел, яка була пов’язана з нашою проблемою. Ми вивчаємо наукові статті, дослідження та інші джерела, щоб зрозуміти, які методи рішень існують у сфері, що нас цікавить.

Після отримання необхідних знань ми перейдемо до вибору підходящого алгоритму, мови програмування та середовища розробки для досягнення нашої

мети. Ми оцінюємо різні алгоритми, їх ефективність і застосовність до нашої задачі. Залежно від характеристик проблеми, обираємо алгоритми машинного навчання, які забезпечувють оптимальний результат.

Після вибору алгоритму визначаємо мову програмування, яка найкраще підходить для реалізації нашого алгоритму, звертаючи увагу на можливості, продуктивність та доступні бібліотеки для машинного навчання. Далі, ми обираємо відповідне середовище для розробки нашого проекту. Ми розгляємо різні інструменти і середовища, які надавають зручний інтерфейс для програмування, можливості для візуалізації даних та відладки. При виборі середовища ми враховувємо також сумісність з обраною мовою програмування.

У кінцевому результаті нашої роботи, ми отримаємо спосіб ідентифікації та кластеризовані молекули за допомогою машинних алгоритмів. Наш підхід дозволяє систематизувати дані та виявити залежності між молекулярними структурами.

РОЗДІЛ 2. АНАЛІЗ СТАНУ ПРОБЛЕМИ

1.1. Машинне навчання в хемо-інформатиці та відкритті ліків.

Машинне навчання на сьогоднішній день є однією з найважливіших та швидко розвиваються галузей у комп'ютерній допомозі у відкритті ліків [1]. У порівнянні з фізичними моделями, які базуються на явних фізичних рівняннях, таких як квантова хімія або симуляції молекулярної динаміки, підходи машинного навчання використовують алгоритми розпізнавання зразків для виявлення математичних зв'язків між емпіричними спостереженнями малих молекул та їхніми хімічними, біологічними та фізичними властивостями, що дозволяє передбачати ці властивості нових сполук. Крім того, порівняно з фізичними моделями, методи машинного навчання є більш ефективними та можуть легко масштабуватись для великих наборів даних без потреби великих обчислювальних ресурсів. Одним з основних напрямків застосування машинного навчання у відкритті ліків є допомога дослідникам у розумінні та використанні зв'язків між хімічною структурою сполук і їхньою біологічною

активністю або SAR [2]. Наприклад, маючи результати скринінгу лікарських речовин, може бути бажано знати, як можна оптимізувати хімічну структуру сполуки для покращення її зв'язування, біологічної активності або фізико-хімічних властивостей. Півстоліття тому такі проблеми можна було вирішити лише шляхом численних дорогих, часоємних і працезатратних циклів синтезу та аналізу лікарської хімії. Сьогодні сучасні методи машинного навчання можуть бути використані для моделювання QSAR, або кількісних відношень структура-властивість (QSPR), і розробки програм штучного інтелекту, які точно передбачають в силіко, як хімічні модифікації можуть вплинути на біологічну поведінку [3]. Багато фізико-хімічних властивостей ліків, таких як токсичність, обмін речовин, міжлікарські взаємодії та канцерогенез, ефективно моделюються методами QSAR [3]. Перші моделі QSAR, такі як аналіз Ханша і Фрі-Вільсона, використовували прості багатовимірні регресійні моделі для кореляції потенції (logIC50) з підструктурними мотивами та хімічними властивостями, такими як розчинність (logP), гідрофобність, субститутний взірець і електронні фактори [4]. Хоча ці підходи були переломними і успішними, вони в кінцевому рахунку обмежувалися недоступністю експериментальних даних та припущенням про лінійність моделювання. Тому необхідні високорівнева хемоінформатика та методи машинного навчання, які здатні моделювати нелінійні набори даних, а також великі дані зі зростаючою глибиною та складністю.

1.2 Кластеризація хімічних сполук

Хімічний аналіз подібності є фундаментальною технікою для пошуку лігандів на основі відкриття ліків [5]. Його мета полягає в ідентифікації та поверненні сполук з бази даних, що мають схожі структури та біоактивність, що подібна до запитових сполук [6]. Принцип хімічної подібності, який стверджує, що сполуки зі схожими структурами, ймовірно, матимуть схожу біоактивність, є основною передумовою подібності заснованого на лігандах віртуального скринінгу [7].

Техніки машинного навчання можна широко класифікувати як навчання з

учителем або без учителя [8]. У випадку навчання з учителем мітки призначаються навчальним даним, і після тренування модель може передбачати мітки для заданих вхідних даних. Моделі машинного навчання з учителем включають регресійний аналіз, метод найближчих сусідів (kNN), Баєсове ймовірнісне навчання, метод опорних векторів (SVM), випадкові ліси та нейронні мережі. Техніки машинного навчання без учителя навчаються безпосередньо на невизначених даних, виявляючи основні закономірності молекулярних ознак. Особливим випадком навчання з учителем є напівнаглядоване навчання або трансдуктивне навчання, при якому невелика кількість позначених даних змішується з непозначеними даними в процесі навчання для покращення точності моделювання невеликого і незбалансованого набору даних [9]. Техніки без учителя включають методи зменшення розмірності, такі як аналіз головних компонентів, аналіз незалежних компонентів, а також деякі методи з учителем, які також підтримують безумовне навчання, такі як методи опорних векторів, ймовірнісні графічні моделі та нейронні мережі [10, 11, 12, 13]. Алгоритми кластеризації є ще одним класом алгоритмів без учителя, де спочатку набір даних розділяється за попередньо визначеними метриками високоінтенсивного простору, а потім мітки присвоюються на основі кількості спостережуваних категорій. Сучасні техніки машинного навчання надають потужний набір методів для дослідження нелінійних зв'язків між хімічною структурою і активністю з високою точністю і прецизією.

Завдяки аналізу даних джерел і враховуючи досвід класифікації в сфері хімінформатики, у нашому дослідженні ми прийшли до висновку, що ефективним способом для вирішення нашої задачі- ідентифікації хлороганічних молекул, буде саме застосування машинного алгоритму. Для цього планується використати метод класифікації “Дерево рішень”, середовище R-Studio та мову R. Але для початку, розбити програму реалізації на 2 алгоритми кластеризації. Перший буде зроблений без застосування машинних методів, в якому буде побудований гіперпростір векторів ознак мас-спектрів. Для швидкої

ідентифікації хлорорганічних молекул,ми побудуємо класифікатор, що складатиметься з 10 класів. Для більшої наочності, кластеризуємо результат методом K-means.

РОЗДІЛ 3. МЕТОДИ ТА МАТЕРІАЛИ

3.1. Мас-спектроскопічний метод аналізу

Метод мас-спектроскопії є способом вивчення речовини шляхом визначення маси іонів, що утворюються з цієї речовини (зазвичай відношення маси іонів до їх заряду) та їх кількості. Сукупність значень мас і відносних вмістів (концентрацій) цих іонів утворює мас-спектр. На рисунку 2.1 представлений вигляд мас-спектра дельтметрину.

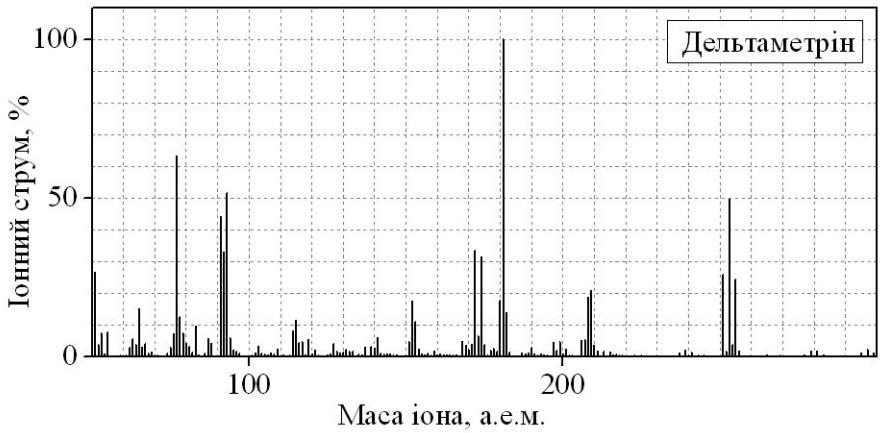

Рис. 1.1. Мас-спектр дельтаметрину

У мас-спектроскопії, речовина піддається іонізації, що полягає у процесі поділу іонів різної маси за впливу електричних і магнітних полів у вакуумі. Однак, при дослідженні іонної структури вже іонізованих газів, наприклад, у електричному розряді або в іоносферах планет, процес іонізації не застосовується. Для рідких і твердих речовин, їх спочатку піддають випаровуванню, а потім іонізують. Частіше вивчають позитивні іони, оскільки існуючі методи іонізації дозволяють отримувати їх простіше і у більших кількостях, ніж негативні іони. Проте, у деяких випадках також досліджують негативні іони.

У Великобританії Дж. Дж. Томсоном у 1910 році та Ф. Астоном у 1919 році були отримані перші мас-спектри, що привели до відкриття стабільних ізотопів. Початково мас-спектроскопію використовували для визначення ізотопного складу елементів та точного виміру їх атомних мас. До сьогодні мас-спектрометрія залишається одним з основних методів отримання даних про маси ядер та атомні маси елементів.

Метод мас-спектрометрії дозволяє визначити варіації ізотопного складу елементів з високою точністю, з погрішністю приблизно ±10-2 % для ізотопного складу та ±10-5 % для мас ядер легких елементів та ±10-4 % для важких елементів. Чутливість і точність мас-спектрометрії призвели до її застосування в інших областях, де знання ізотопного складу елементів має важливе значення, зокрема в ядерній техніці.

У геології та геохімії мас-спектральний аналіз ізотопного складу різних елементів, таких як свинець чи аргон, є основою для визначення віку гірських порід та рудних утворень. В хімії мас-спектрометрія широко використовується для елементного та молекулярного структурного аналізу.

Молекулярний мас-спектральний аналіз речовин особливо точний, коли речовина випаровується у вигляді недисоційованих молекул, які не розпадаються в іонному джерелі спектрометра. Застосовуючи мас-спектрометри з високою роздільною здатністю, можна визначити число атомів вуглецю, водню, кисню та інших елементів у молекулі органічної речовини за масою молекулярного іону.

Метод іонізації вакуумною іскрою застосовується для аналізу елементного складу nebolatile речовин. Якісний та кількісний молекулярний мас-спектральний аналіз сумішей базується на різниці мас-спектрів молекул різної будови та пропорційності іонних струмів компонентів суміші до їх вмісту.

В кількісному молекулярному аналізі точність, в найкращому випадку, може досягати рівня точності ізотопного аналізу. Однак, у багатьох випадках кількісний молекулярний аналіз ускладнюється через рівність мас різних іонів,

що утворюються при іонізації різних речовин. Для подолання цих труднощів у мас-спектрометрах використовуються "м'які" методи іонізації, які породжують меншу кількість фрагментованих іонів. Також поширені комбінації мас-спектрометрії з іншими методами аналізу, зокрема з газовою хроматографією, для подальшого підвищення точності та надійності результатів.

Молекулярний структурний мас-спектральний аналіз базується на тому, що під час іонізації речовини деяка частина молекул перетворюється на іони без руйнування, тоді як інша частина розпадається на фрагменти. Вимірювання мас і відносного вмісту молекулярних і осколкових іонів (молекулярний мас-спектр) надає інформацію не лише про молекулярну масу, але й про структуру молекули.

Мас-спектрометрія широко використовується в фізико-хімічних дослідженнях для вивчення процесів іонізації та дослідження фізичної й хімічної кінетики. Вона також застосовується для визначення потенціалів іонізації, теплоти випаровування та енергії зв'язку атомів у молекулах. Мас-спектрометрія була використана для вимірів складу атмосфери Землі, аналогічні виміри проводяться і для атмосфер інших планет. У медицині вона знаходить застосування як експресний метод газового аналізу. Ще однією перевагою мас-спектрометрії є її висока абсолютна чутливість, що дозволяє аналізувати надзвичайно малі кількості речовини (на рівні 10-12 грама).

Теоретичні основи мас-спектроскопії

У мас-спектрометрії досліджують питомі заряди іонів, які визначаються відношенням заряду частинки до її маси (q/m). Для вимірювання питомого заряду використовують відхилення заряджених частинок у магнітному та електричному полях. Радіус траєкторії частинки обчислюється за допомогою формули: r = mV/qB. Для виявлення іонів використовують фотопластинки та детектори іонів.

Теорія молекулярного структурного мас-спектрального аналізу, яка базується на методі іонізації електронним ударом (з використанням електронів з високою енергією), передбачає утворення збудженого молекулярного іона

після удару, який потім розпадається зі зламом слабких зв'язків у молекулі. Зараз теорія ще не може передбачити точні мас-спектри молекул і не надає необхідних коефіцієнтів чутливості для кількісного аналізу різних речовин. Тому для визначення невідомої структури молекули за мас-спектром і для якісного аналізу використовують кореляційні дані між мас-спектрами речовин різних класів. Для оцінки коефіцієнта чутливості використовують лінійний зв'язок між сумарною ймовірністю іонізації та молекулярною масою для нескладних молекул одного гомологічного ряду. Тому при молекулярному мас-спектральному аналізі, якщо це можливо, проводять градуювання приладу за допомогою відомих речовин або сумішей відомих сполук.

Сучасні мас-спектрометри використовують комп'ютерну техніку для аналізу та інтерпретації отриманих даних. Вбудовані процесори використовуються для обробки даних мас-спектроскопічного аналізу та його результатів.

3.2. Кластеризація та класифікація

Кластерний аналіз, також відомий як аналіз скупчень або групування даних, є задачею поділу заданої вибірки об'єктів (ситуацій) на підмножини, які називаються кластерами. Метою кластерного аналізу є створення кластерів, в яких об'єкти внутрішньо подібні між собою, тоді як об'єкти з різних кластерів суттєво відрізняються один від одного. У цьому випадку термін "кластер" походить від англійського слова "cluster", що означає "гроно", "об'єднання" або "скупчення".

Під час кластеризації спостережень головною метою є неформальне та економічне обґрунтування поділу певної сукупності на групи. Кластерний аналіз дозволяє проводити цей поділ, не вимагаючи заздалегідь встановлених значень про розподіл генеральної сукупності. Кожен об'єкт може бути описаний набором спостережуваних параметрів, які можна розглядати як координати точок у багатомірному просторі.

Основною метою кластерного аналізу є поділ об'єктів на групи, які

виявляють відносну однорідність або схожість за досліджуваними ознаками або змінними. Кожна група містить об'єкти, які подібні між собою за досліджуваними показниками, в той час як об'єкти в різних групах суттєво відрізняються. При використанні кластерного аналізу шляхом групування в меншу кількість кластерів, зменшується кількість об'єктів у кожному кластері, але не зменшується кількість досліджуваних ознак чи змінних.

Методи кластерного аналізу можуть бути застосовані в різних ситуаціях, включаючи просте групування об'єктів на основі їх подібності за кількісними ознаками. Відмінною особливістю кластерного аналізу є те, що він не є типовим статистичним методом, і в багатьох випадках не вимагає процесів статистичної перевірки значущості. Це дозволяє використовувати кластерний аналіз, наприклад, при статистичному аналізі даних, векторній квантизації, розпізнаванні образів та інших задачах. Однією з переваг кластерного аналізу є можливість розбивати об'єкти на групи, враховуючи не одну ознаку, а цілий набір ознак. Крім того, кластерний аналіз не накладає обмежень на вигляд даних і дозволяє використовувати різноманітні дані різної природи. Це особливо корисно, наприклад, при прогнозуванні економічних показників, коли вхідні дані мають різну структуру і традиційні економетричні підходи стають непридатними. Задача кластерного аналізу може бути сформульована так: заданий набір з n векторів, кожен з яких має d ознак; метою є розбиття на підмножини відповідно до заданого критерію оптимізації. Зазвичай, цей критерій полягає у мінімізації спотворення. Існують різні методи оцінювання спотворення, але у більшості прикладних реалізацій використовується сума середньо-квадратичних відстаней Евкліда між вектором і центроїдом кластера, до якого він належить.

Перш ніж перейти до конкретних методів та алгоритмів кластеризації, важливо врахувати кілька ключових моментів:

- Кластерний аналіз використовується для виявлення структури в даних, але водночас він може впроваджувати структуру, якої насправді немає. Це означає, що результати кластерного аналізу можуть бути

артефактами штучної структури, якої насправді немає.

- Більшість методів кластерного аналізу є евристичними процедурами і не мають строгого статистичного обґрунтування. Вони базуються на певних припущеннях та спрощеннях, що можуть впливати на точність і надійність отриманих результатів.
- Різні методи кластеризації можуть призводити до різних кластерних розбиттів для тих самих даних. Це нормальне явище, і вибір конкретного методу кластеризації повинен бути обґрунтованим і документованим.
- Для здійснення осмисленого вибору при кластерному аналізі потрібна теоретична база або гіпотеза про структуру даних. Без такої теоретичної моделі існує ризик безкритичного емпіризму, коли результати кластерного аналізу приймаються без обґрунтування. Кластерний аналіз повинен підкріплюватись науковими основами.

Варто пам'ятати, що кластерний аналіз, як і будь-який інший метод, має свої обмеження та недоліки.

Для проведення кластерного аналізу необхідно враховувати наступні етапи:

- Визначення характеристик, за якими будуть оцінюватися об'єкти в досліджуваній вибірці.
- Вибір оптимальної кількості кластерів, до яких будуть групуватися об'єкти.
- Обчислення міри схожості між об'єктами для визначення ступеня їхньої подібності.
- Обґрунтування вибору певного методу або алгоритму кластерного аналізу для створення груп схожих об'єктів.
- Перевірка достовірності результатів кластеризації шляхом застосування відповідних перевірочних процедур.
- Подання та інтерпретація отриманих результатів, що включає пояснення сутності кожного кластера та його значення.

При проведенні кластерного аналізу можуть виникати деякі проблеми, такі як перекручення даних при зведенні їх до компактного вигляду або втрата індивідуальних рис об'єктів через узагальнення їхніх характеристик в межах кластеру. Також варто пам'ятати, що в деяких випадках можуть бути виявлені порожні кластери, коли деякі значення не належать жодному кластеру.

Важливо мати теоретичну базу або гіпотезу про структуру даних, щоб зробити осмислені вибори при кластерному аналізі. Кластерний аналіз має свої обмеження та недоліки, які потрібно враховувати при його застосуванні.

Класифікація - це систематичне розподілення об'єктів, явищ, процесів на роди, види, типи з метою зручного вивчення. Цей процес включає угруповання понять і їх розташування у відповідному порядку, що відображає ступінь схожості. Класифікація полягає в упорядкуванні об'єктів залежно від спільних ознак або властивостей, які використовуються для визначення подібності або відмінності між ними. Класифікація дозволяє зробити висновки про характеристики конкретної групи. Цей процес ґрунтується на наявності ознак, що характеризують групу, до якої належить об'єкт або подія. Класифікація також використовується в стратегії навчання з учителем, коли вже існують класифіковані об'єкти для формулювання правил. Патерни спектрів, накопичені в базі даних, використовуються як для ідентифікації компонентів спектра, так і для створення систем рівнянь, що дають точні рішення задачі.

3.4. Найпоширеніші методи класифікації

3.4.1. Класифікатор K-means

K-середніх (англ. K-means) - це популярний алгоритм кластеризації, який використовується в машинному навчанні та добуванні даних. Це некерований алгоритм навчання, який групує схожі точки даних разом на основі їх схожості за ознаками.

K-середніх є ітеративним алгоритмом кластеризації, який можна поділити на кілька кроків. Початкові центроїди кластерів можуть бути ініціалізовані випадковим чином або вибрані з деякими попередніми знаннями про дані.

У кроці призначення кожна точка даних обчислює відстань до всіх центроїдів і призначається до кластера з найближчим центроїдом. Відстань може бути обчислена, наприклад, за допомогою Евклідової відстані або косинусної схожості, залежно від характеру даних і вимог задачі.

Після призначення всіх точок даних до кластерів, переходимо до кроку оновлення центроїдів. Нові центроїди кластерів обчислюються шляхом визначення середнього значення всіх точок у кожному кластері. Це означає, що центроїди переміщуються до нових позицій залежно від розташування точок у кластерах.

Процес призначення та оновлення центроїдів повторюється до досягнення збігу або задоволення умови зупинки. Збіг означає, що центроїди не змінюють своє положення між ітераціями, тобто кластеризація зберігається. Умова зупинки може бути встановлена на основі максимальної кількості ітерацій або зміни внутрішньокластерних дисперсій.

Одна з головних переваг алгоритму K-середніх полягає в його ефективності, особливо при великих обсягах даних. Оскільки алгоритм є некерованим, він може автоматично виявляти структуру даних без попередніх знань про класи або кластери. Крім того, K-середніх є відносно простим і легким для реалізації алгоритмом.

Оптимізації та варіанти алгоритму K-середніх розробляються для покращення його продуктивності та якості кластеризації. Наприклад, існують підходи до вибору оптимального значення K, такі як метод ліктя або метод силуету. Існують також розширені версії алгоритму, які уникнуть недоліків базового алгоритму, такі як K-середніх++, який поліпшує початкову ініціалізацію центроїдів, або K-медоїдів, який використовує медоїди замість середніх значень.

Алгоритм K-середніх має широке застосування в різних областях, включаючи сегментацію зображень, аналіз даних, групування користувачів, рекомендації, біоінформатику та інші. Він дозволяє виявляти структуру та групи у наборах даних, що допомагає зрозуміти характеристики даних та

здійснювати подальший аналіз та вивчення.

3.4.2. Класифікатор SVM

Метод опорних векторів (Support Vector Machines, SVM) є одним з методів, який використовується для класифікації та регресії. Він базується на побудові нелінійної границі прийняття рішень у просторі ознак.

Основна ідея методу полягає у знаходженні оптимальної гіперплощини, яка розділяє два класи об'єктів у просторі ознак. Ця гіперплощина вибирається таким чином, щоб максимізувати відстань (заздалегідь визначену як "зазор") між найближчими об'єктами двох класів, які знаходяться поруч з границею розділення. Ці об'єкти, які лежать найближче до границі, називаються опорними векторами.

Метод опорних векторів має високу гнучкість і може вирішувати задачі класифікації та регресії навіть у випадку, коли дані не є лінійно роздільними.

Під час процесу навчання метод опорних векторів шукає оптимальні коефіцієнти, що визначають положення та орієнтацію гіперплощини, так щоб вона максимально відділяла точки класів. Ці коефіцієнти враховуються для побудови роздільної гіперплощини, а також для ідентифікації опорних векторів - точок, що знаходяться найближче до гіперплощини.

3.4.3. Класифікатор Naive Bayes

Байєсовський класифікатор є широким класом алгоритмів класифікації, які базуються на принципі максимуму апостеріорної ймовірності. Цей підхід використовується для визначення приналежності класу для заданого об'єкта шляхом обчислення функцій правдоподібності для кожного класу і наступного обчислення апостеріорних ймовірностей для кожного класу. Об'єкт відноситься до класу, для якого апостеріорна ймовірність є найбільшою.

Наївна класифікація є досить зрозумілим і прозорим методом класифікації. Вона отримала назву "наївна" через припущення про взаємну незалежність ознак. Властивості наївної класифікації включають використання всіх змінних і визначення всіх залежностей між ними. При цьому метод базується на двох припущеннях щодо змінних: всі змінні мають однакову вагу і

є статистично незалежними, що означає, що значення однієї змінної не надає жодної інформації про значення іншої змінної.

Більшість інших методів класифікації припускають, що ймовірність належності об'єкта до класу є однаковою для всіх класів перед початком класифікації. Проте це не завжди є правильним припущенням.

Дані про відсоток даних, що належать конкретному класу, можна використовувати при побудові моделі класифікації. Ця апріорна інформація може бути використана як додаткові дані під час побудови класифікаційної моделі.

Отже, використання байєсовської класифікації та байєсовських мереж дозволяє використовувати апріорні знання та додаткові дані для покращення процесу класифікації.

Наївно-байєсовський підхід, як і будь-який інший метод, має свої недоліки. Декілька недоліків наївно-байєсовського підходу включають наступне:

- Припущення про статистичну незалежність змінних: У наївному-байєсовському підході передбачається, що всі вхідні змінні є статистично незалежними. У реальних даних це припущення часто не виконується, але метод може все одно показувати прийнятні результати. Проте, при недотриманні умови статистичної незалежності, більш складні методи, такі як байєсовські мережі з навчанням, можуть бути більш ефективними.
- Обробка безперервних змінних: Наївно-байєсовський підхід не може безпосередньо обробляти безперервні змінні. Для використання цих змінних, вони повинні бути перетворені на дискретну шкалу, наприклад, розбиваючи їх на інтервали. Проте, такі перетворення можуть призводити до втрати значимих закономірностей у даних.

Враховуючи ці недоліки, важливо усвідомлювати, що наївно-байєсовський підхід є простим інструментом, який може бути ефективним у деяких випадках, але не завжди відповідає всім особливостям даних. При роботі з складнішими даними і при наявності більш точних методів, може бути

доцільним розглядати інші підходи, які дозволять уникнути обмежень наївного-байєсовського підходу.

3.4.4. K-найближчих сусідів(k-NN)

Алгоритм K-найближчих сусідів (k-NN) є одним з керованих алгоритмів машинного навчання, який може використовуватися як для класифікації, так і для задач прогнозування регресії. Основне застосування KNN полягає в класифікації проблем у промисловості. k-NN можна вважати алгоритмом ледачого навчання, оскільки він не має окремого етапу навчання та використовує всі наявні дані для класифікації. Крім того, k-NN є непараметричним алгоритмом навчання, що означає, що він не передбачає жодних певних припущень про базові дані або їх розподіл.

Алгоритм K-найближчих сусідів (k-NN) використовує "схожість ознак" для прогнозування значень нових точок даних. Це означає, що нова точка даних отримує значення на основі того, наскільки вона близька до інших точок у навчальному наборі. Проте, алгоритм k-NN має кілька недоліків, серед яких можна виділити наступні:

Швидкість роботи: В реальних задачах зазвичай потрібно використовувати велику кількість сусідів для класифікації (наприклад, 100-150). У такому випадку алгоритм може працювати повільніше, ніж, наприклад, дерева рішень.

Великий простір ознак: Якщо простір ознак об'єкта є великим, важко підібрати вагові коефіцієнти та визначити, які ознаки є неважливими для вирішення задачі.

Залежність від метрики відстані: Вибір певної метрики виміру відстані між об'єктами може суттєво вплинути на результат. Використання за замовчуванням евклідової відстані не завжди є обґрунтованим, і необхідно ретельно вибирати відповідну метрику. Однак, пошук оптимальних параметрів може бути часомоємним для великого обсягу даних.

Вибір кількості сусідів: Немає теоретичного обґрунтування для вибору кількості сусідів. Часто використовується перебір параметрів, що може бути

часомоємним. Крім того, недостатня кількість сусідів може призводити до перенавчання і зробити метод чутливим до викидів.

При виборі кількості сусідів k для алгоритму k-NN необхідно уникати крайніх значень, так як вони можуть призвести до неправильних класифікацій. Коли k = 1, алгоритм найближчого сусіда стає нестійким до шумових викидів, і може неправильно класифікувати не лише самі викиди, але й найближчі до них об'єкти інших класів. З іншого боку, при k = m, алгоритм стає надмірно стійким і перетворюється на константу, що не дає корисних результатів.

На практиці оптимальне значення параметра k визначають за допомогою критерію змінного контролю, часто застосовуючи метод виключення об'єктів по одному (leave-one-out cross-validation). Цей метод дозволяє оцінити ефективність алгоритму для різних значень k, виключаючи по одному об'єкту з навчального набору і тестуючи його на залишкових даних. Значення k, яке надає найкращу точність чи інший метричний показник, вважається оптимальним для даної задачі класифікації.

3.4.5. Класифікатор Random Forest

Random Forest є ансамблем рішень, що складається з багатьох вирішувальних дерев. Цей метод дозволяє зменшити проблему перенавчання та підвищити точність порівняно з окремим деревом. Прогноз формується шляхом комбінування відповідей декількох дерев. Кожне дерево навчається незалежно від інших на різних підмножинах даних, що уникне побудови однакових дерев на одних і тих самих даних. Це робить алгоритм ефективним для використання в розподілених системах обчислень. Ідея беггінгу, запропонована Лео Брейманом, є доречною для розподіленого обчислення.

Під час беггінгу, що використовується для незалежного навчання класифікаційних алгоритмів, розумно використовувати багато дерев рішень зі значною глибиною. Під час класифікації, остаточним результатом буде клас, за яким проголосувало більшість дерев, при умові, що кожне дерево має один голос.

Наприклад, у випадку бінарної класифікації з використанням моделі з 500

дерев, які голосують, 100 з них вказують на перший клас, а решта 400 - на другий клас, модель буде передбачати другий клас. У випадку застосування Random Forest для задач регресії, підхід вибору значення, за яким проголосувала більшість дерев, буде непридатним. Замість цього обирається середнє значення, отримане з усіх дерев.

Random Forest потребує значних обчислювальних ресурсів через незалежне будування глибоких дерев. Обмеження на глибину може погіршити точність, оскільки для складних задач необхідно побудувати багато глибоких дерев. Час навчання дерев збільшується лінійно з їх кількістю.

Зрозуміло, що збільшення глибини дерев не найкращим чином впливає на продуктивність, але підвищує ефективність алгоритму (хоча збільшує ризик перенавчання). Важливо не переживати занадто через перенавчання, оскільки це компенсується кількістю дерев. Але також не слід піддаватися перенавантаженню. У всіх випадках важливо правильно налаштувати гіперпараметри.

3.4.6. Машинний алгоритм «Дерево рішень»

Дерево рішень може використовуватися як класифікатор у машинному навчанні. Дерево рішень - це модель машинного навчання, яка використовується для класифікації або регресії даних. У випадку класифікації, дерево рішень розділяє набір даних на декілька підмножин, основуючись на значеннях вхідних ознак, і призначає кожній підмножині певний клас або мітку. Таким чином, дерево рішень може використовуватися як класифікатор для призначення міток або класів новим невідомим даним, основуючись на вивчених зразках даних. Алгоритм дерева рішень побудований у вигляді дерева з вузлами та гілками. Кожен вузол представляє тест на певну ознаку даних, а кожна гілка відображає можливі результати цього тесту. Вузли дерева рішень розподіляють дані на підмножини залежно від значень ознаки, і рішення приймається на основі значення цільової змінної в листках дерева.

Дерева рішень мають деякі переваги, такі як легка інтерпретація, здатність враховувати неоднорідність даних та можливість роботи з якісними та

кількісними ознаками. Однак вони також можуть бути схильні до перенавчання, особливо якщо дерево має велику глибину та багато правил.

Загалом, дерева рішень є одним із широко використовуваних алгоритмів машинного навчання та знаходять своє застосування в багатьох галузях.

3.5. Середовище розробки R-Studio та платформа R

R-Studio - це інтегроване середовище розробки (IDE) для мови програмування R. R-Studio надає зручний і потужний інтерфейс для роботи з мовою R, дозволяючи програмістам і дослідникам швидко створювати, тестувати і відлагоджувати свої програм.

Мова програмування R є високорівневою мовою програмування, розробленою спеціально для аналізу даних і статистичних обчислень. Вона надає багатий набір функцій і пакетів для статистичного аналізу, візуалізації даних, машинного навчання та інших завдань аналітики даних.

R має синтаксис, подібний до мови програмування S, і надає велику кількість вбудованих функцій і операторів для роботи з даними. Вона підтримує векторні операції, що дозволяють ефективно обробляти масиви даних. Багато користувачів R також використовують його для візуалізації даних за допомогою графіків і діаграм.

R-Studio надає зручні інструменти для написання, відлагодження і виконання програм на мові R, а також має можливості для роботи з проектами, керування пакетами, керування історією команд та багато іншого. Воно стало популярним серед аналітиків даних і дослідників через свою потужну функціональність і дружній інтерфейс.

Використовуючи R-Studio та мову програмування R, можна проводити широкий спектр аналізу даних, від простих статистичних обчислень до складних моделей машинного навчання.

Дерево рішень є одним з методів машинного навчання, який може використовуватися для ідентифікації молекул. Переваги дерев рішень порівняно з іншими класифікаторами, такими як SVM (метод опорних

векторів), k-NN (метод найближчих сусідів), Байєсовський класифікатор та Random Forest. Інтерпретованість: Дерева рішень зазвичай мають просту інтерпретовану структуру, що дозволяє легше розуміти, як саме приймаються рішення. Це важливо в біомедичних та хімічних дослідженнях, де важливо мати зрозумілі та переконливі пояснення. Обробка великих наборів даних: Дерева рішень можуть ефективно опрацьовувати великі об'єми даних, оскільки розділяють простір можливих варіантів на піддерева. Це може бути корисно для ідентифікації молекул у великих базах даних або даних з високою розмірністю. Толерантність до пропущених даних: Дерева рішень можуть працювати з наборами даних, в яких є пропущені значення. Вони можуть автоматично вирішувати, як використовувати доступні атрибути для класифікації, не вимагаючи додаткової обробки даних. Врахування несиметричності класів: Дерева рішень можуть працювати добре з наборами даних, в яких класи мають різну кількість екземплярів. Вони можуть виявляти різні варіанти розділення простору класів, що дозволяє краще моделювати несиметричність.

Отже, за допомогою алгоритму "дерева рішень" було встановлено, що наша побудована модель класифікації дозволяє ідентифікувати невідому молекулу і визначити, до якого класу вона належить. Це досягається шляхом введення нової молекули з її мас-спектральними характеристиками до навченої моделі і аналізу вагомості змінних, використовуючи класифікацію, яка представлена у вигляді графу.

Таким чином, за допомогою моделі класифікації, можна зробити висновок, що невідома молекула є фрагментом молекулярної структури вже відомим сполукам (класам). Цей підхід дозволяє нам ефективно вирішувати задачу ідентифікації молекул і класифікації їх до відповідних класів та показує, що молекула має спільні фрагменти молекулярної структури, що сприяє нашому дослідженню у сфері хлорорганічної хімії.

Додатково, для наглядності ідентифікації молекул, ми можемо скористатись методом кластеризації за допомогою алгоритму k-means. Після

застосування цього методу, кожен окремий кластер можна позначити окремим кольором, що відповідає конкретному класу або сполуці.

Таким чином, кожен кольоровий клас у візуалізації показує групу молекул, які мають схожі піки інтенсивності мас-спектрів та відносяться до одного й того ж класу. Це дозволяє нам легко спостерігати групування молекул і визначати їх класифікацію наочно.

Такий підхід до кластеризації через k-means допомагає візуалізувати результати ідентифікації молекул та зробити їх більш зрозумілими та доступними для аналізу.

РОЗДІЛ 4. РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТУ ТА ОБГОВОРЕННЯ

4.1. Простір ознак хлорогранічних молекул.

Маємо набір даних, мас-спектри хлорогранічних молекул, що складається з 450 вимірювань, у яких є дві ознаки - маса іона і інтенсивність відповідного піка. За координати вектора ознак візьмемо інтенсивність піків. Знаходимо координати радіус-вектора (вектора ознак). Вони будуть дорівнювати інтенсивності, тобто наш радіус-вектор буде мати 450 координат. Далі із експериментальних даних вибираємо опорних радіус-вектор Ro (відносно якого будемо будемо порівнювати інтенсивності спектрів) в нашому випадку це $CH_6H_6C_{16}$ (Ліндан).

Вибираємо мас-спектр який ми будемо порівнювати з спектром CH6H6CI6 (Ліндан). В нашому випадку – це C12H9CI(3-Chlorobiphenyl). Знаходимо його радіус-вектор Rx, тобто беремо координати інтенсивності. Щоб порівняти ці два вектори треба їх відняти Ro-Rx = Rd (Rd – вектор різниці). Знаходимо довжину вектора Rd або просто його модуль.

Також щоб порівняти ці спектри потрібно знайти кут між векторами Ro і Rd. Cos(α) знаходиться з формули 3.1, а кут з 3.2.

$$\cos = Ro*RdRo*Rx \quad (3.1)$$

α=arccos(Ro*RdRo*Rd) (3.2)

Ці кроки повторюємо і для інших мас-спектрів.

Отримані вектори зображаємо в полярній системі координат (рис.3.1)

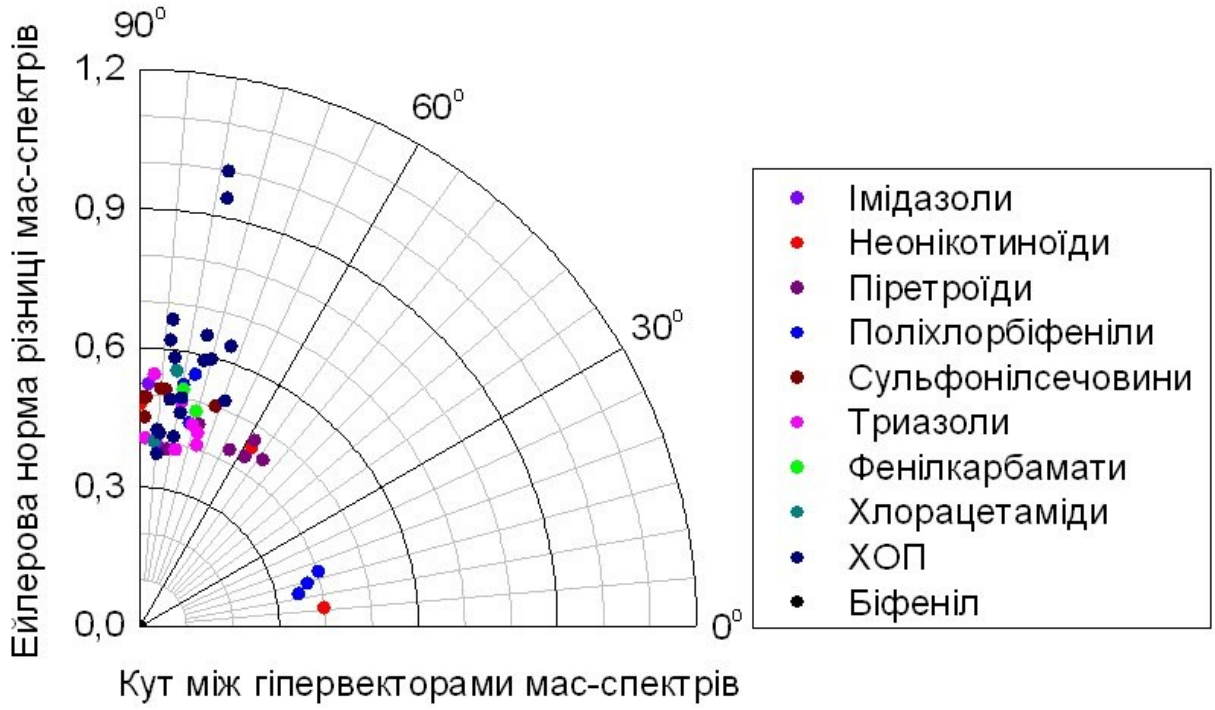

Рис. 4.1. Отримані кластери в полярній системі координат

4.2 Класифікація застосуванням моделі «Дерево рішень»

Були надані 1059 спостережень і 457 змінних. Для класифікації та подальшого прогнозування класу молекул був використаний метод “Дерево рішень”.

Після побудови моделі “Дерево рішень”, багато змінних мали важливість, яка дорівнює нулю, тому вони не впливають на модель. В додатку 2 і на рис. 3.2 наведено змінні, які мають не нульову важливість.

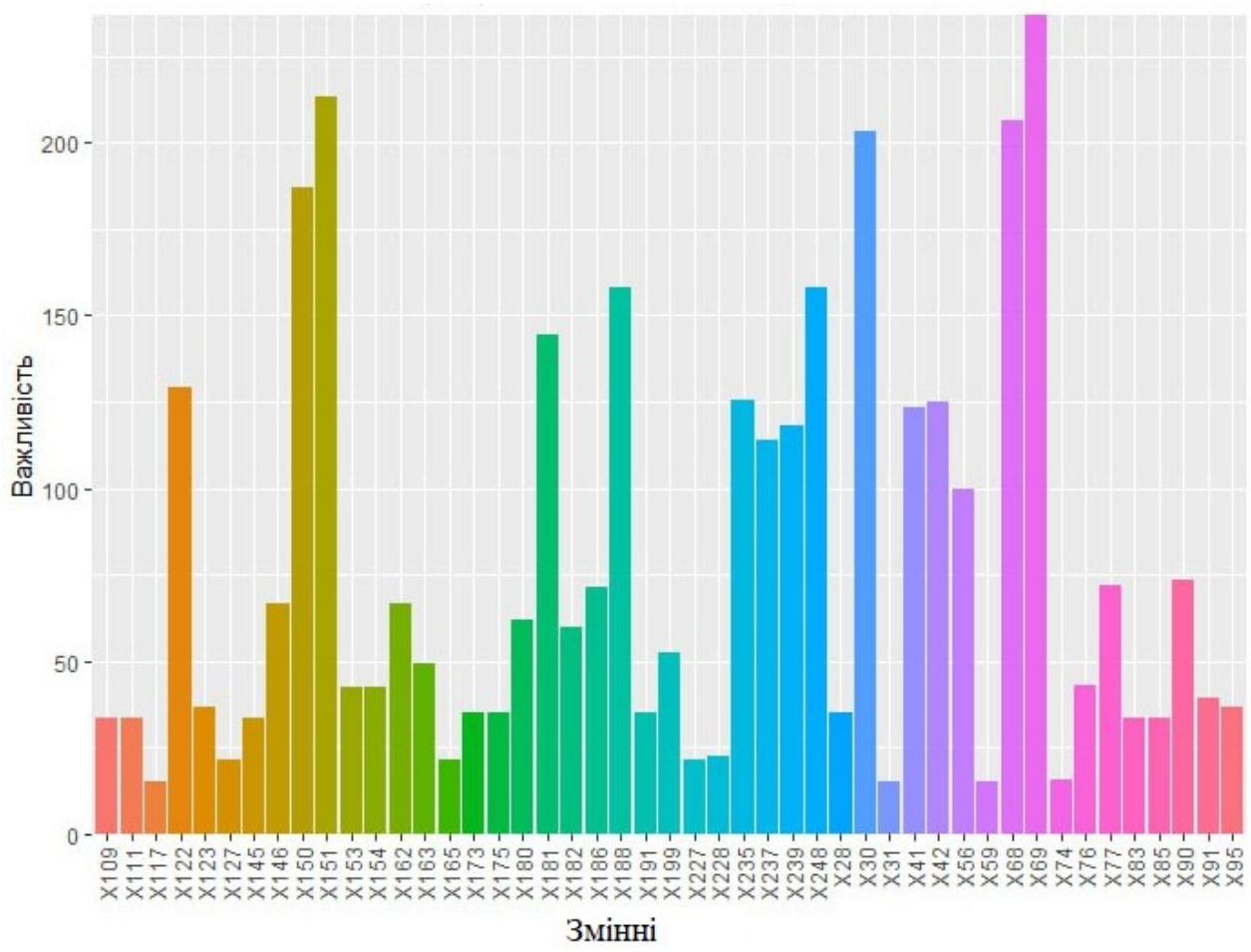

Рис. 4.2. Графік важливості змінних

Вибірка була розбита на навчальну та тестову. Навчальна містить 1000 спостережень, а тестова – 59. Дана модель була побудована на базі тренувальної вибірки і спрогнозована на тестову.

Тренувальна вибірка використовується для навчання моделі. Це підмножина даних, на якій модель будує свої внутрішні правила та параметри. Модель адаптується до тренувальної вибірки шляхом використання алгоритму навчання, щоб встановити оптимальні значення параметрів.

Тестова вибірка, з іншого боку, використовується для оцінки прогностичної здатності моделі. Це незалежна вибірка, яка не використовується під час навчання моделі. На тестовій вибірці ми перевіряємо, наскільки добре модель здатна узагальнити свої знання і робити прогнози на нових, раніше невиданих даних. За допомогою тестової вибірки ми можемо оцінити точність,

чутливість або інші метрики ефективності моделі.

Точність даної моделі дорівнює 100%

На базі отриманих даних було побудовано граф, який зображений на рис. 4.3.

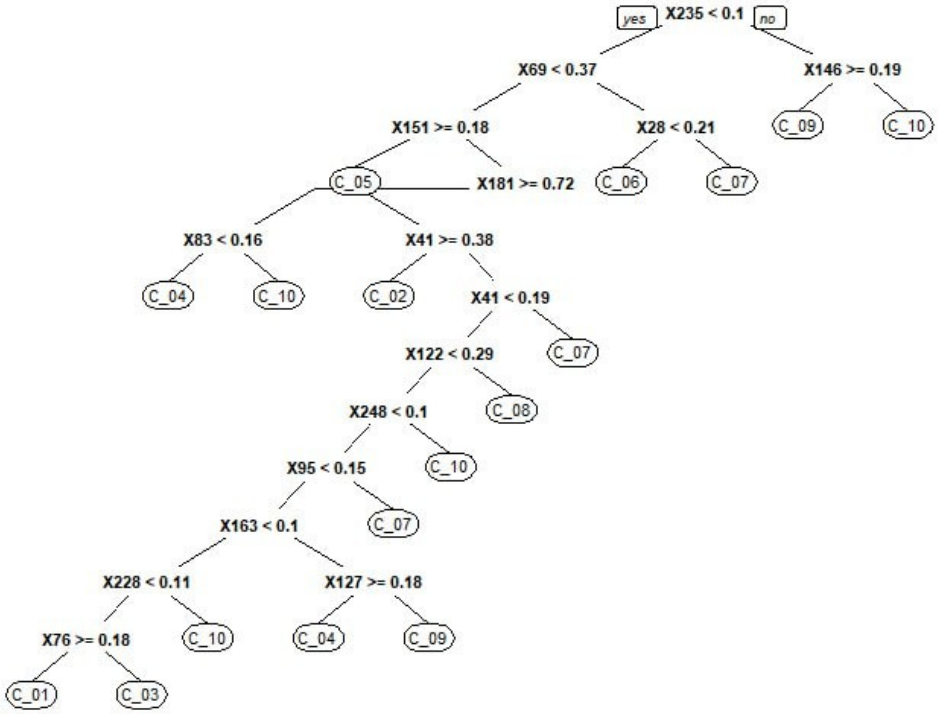

Рис. 4.3. Граф моделі “Дерево рішень”

4.3. Кластеризація хлорогранічних молекул методом k-means

На рис.3.4 зображено кластери, які були отримані методом k-means.

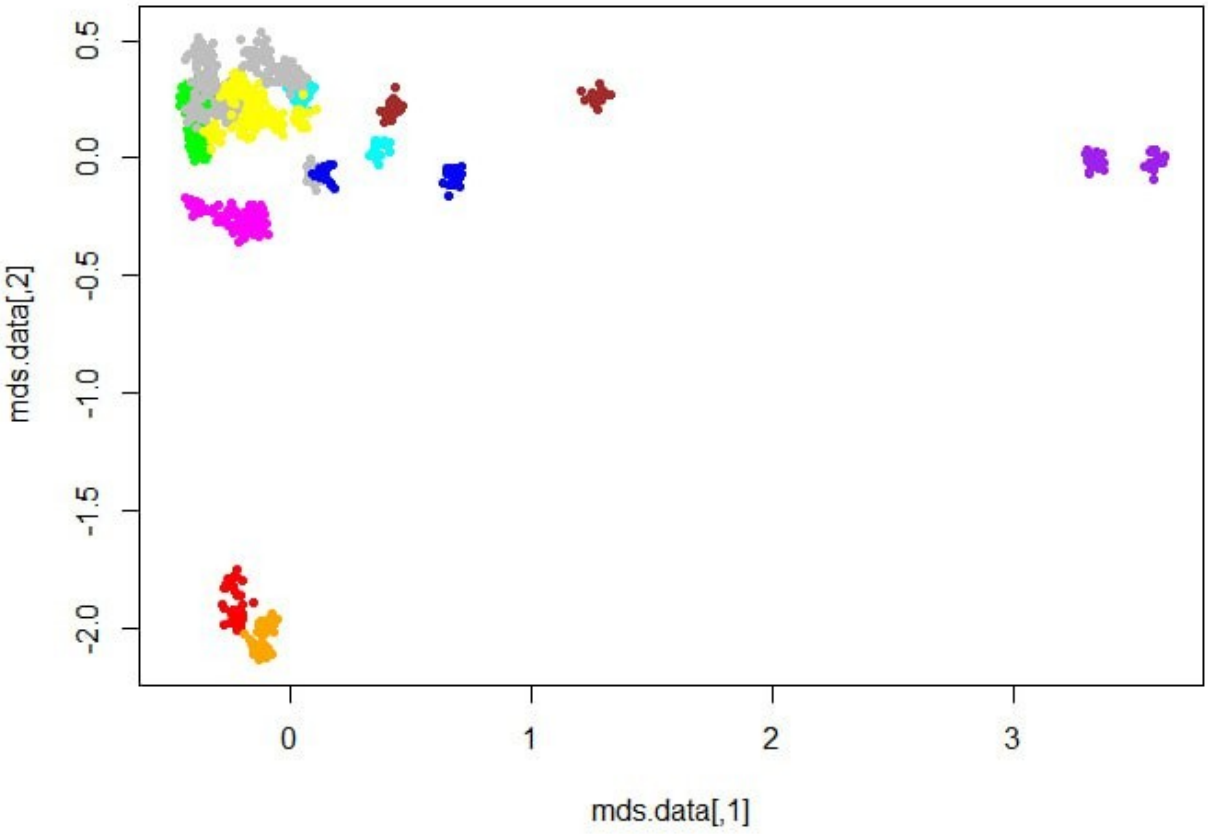

Рис. 4.4. Отримані кластери методом k-means в програмі RStudio

Де, "red" - біфініл "blue" - імідазол, "green" - неонікотиноїд, "orange" - піретроїд ,"purple" - поліхлорбіфеніл, "yellow" - сульфонілсечовина, "cyan" - триазол , "magenta" - фенілкарбамат, "brown" - хлорацетамід, "gray" - ХОП

Цей метод було використано для кластеризації даних в середовищі RStudio. Він є одним з найпоширеніших і простих алгоритмів кластеризації, який дозволяє групувати схожі об'єкти на основі їхніх характеристик. Всі 457 змінних накладаються в двовимірний простір, а перед цим вони шкалуються (наприклад від 1 до 10), щоб не було різких викидів, тобто щоб вони були в

одній розмірності.

На рис. 3.5 зображено графік ліктя.

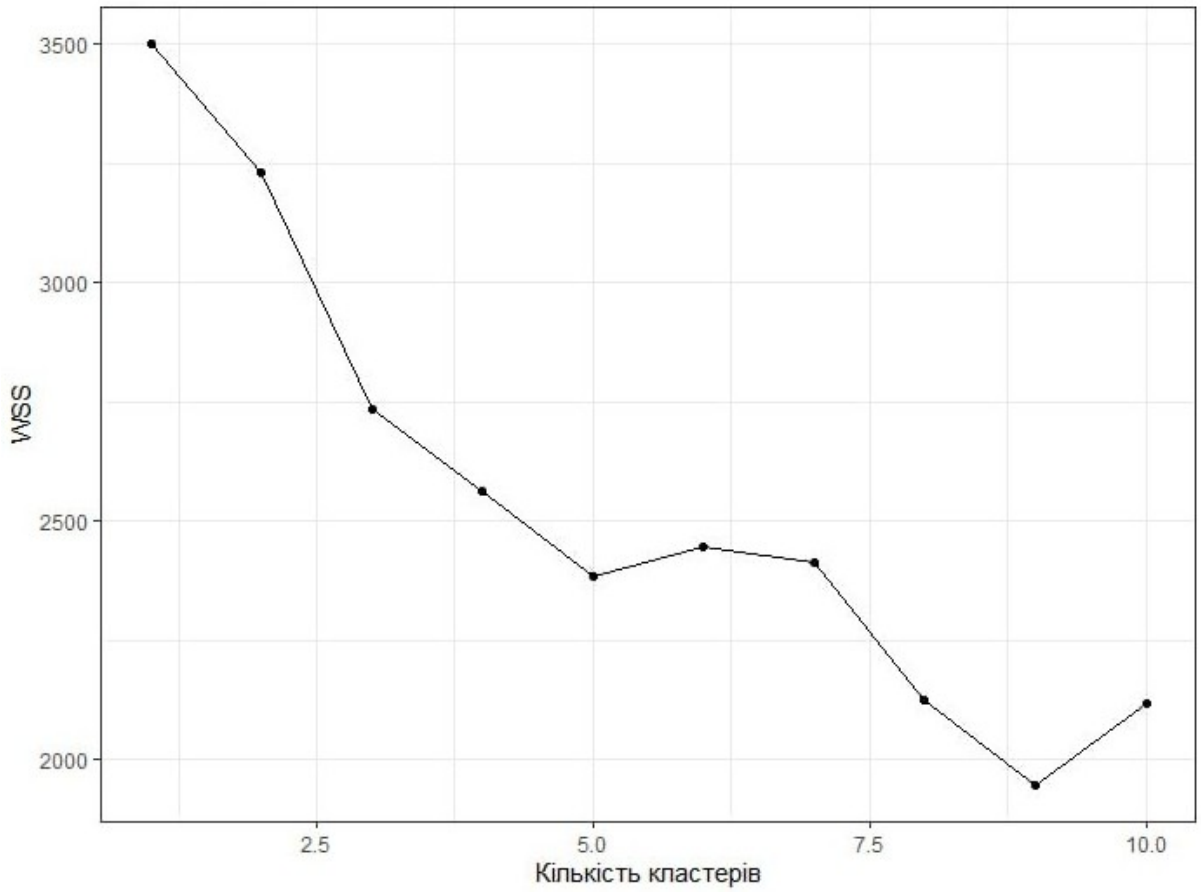

Рис. 4.5. Графік ліктя

Його використовують для визначення кількості кластерів. На графіку локтя по осі абцис відображається кількість кластерів, а по осі ординат - сума внутрішньокластерної дисперсії або сума квадратів відстаней між прикладами даних і центрами їхніх відповідних кластерів (WSS). WSS – це характер змінни дисперсії. Він допомагає визначити, при якій кількості кластерів алгоритм забезпечує найкращу апроксимацію даних. Коли WSS проходить плавно по графіку, то це означає, що дані входять в один кластер. Чим плавніше в нас графік, тим менше кластерів

Таким чином, на графіку локтя оптимальна кількість кластерів відповідає точці, де зменшення внутрішньокластерної дисперсії (або відстані) стає менш помітним (якби лікоть на лікці) після згину. Ця кількість кластерів може бути вибрана як оптимальна для подальшого використання в алгоритмі

кластеризації.

Для прикладу застосування реалізованого класифікатора візьмемо три відомі сполуки, як Цифлутрин, Біфентрин, Дельтаметрин, які відносяться до класу піретроїдів. Також на рис. 4.6, 4.7, 4.8. зображуємо молекулярні структури цих сполук відповідно. Як ми можемо побачити, всі три зображення мають однаковий фрагмент, карбоксильну групу, фенольне кільце, це і є причиною чому вони відносяться до одного класу сполук.

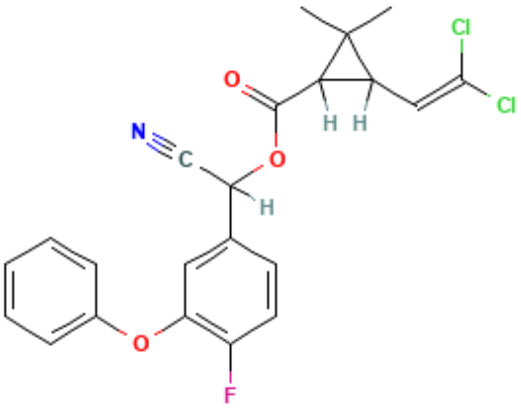

рис. 4.6. Молекулярна структура Цифлутрину

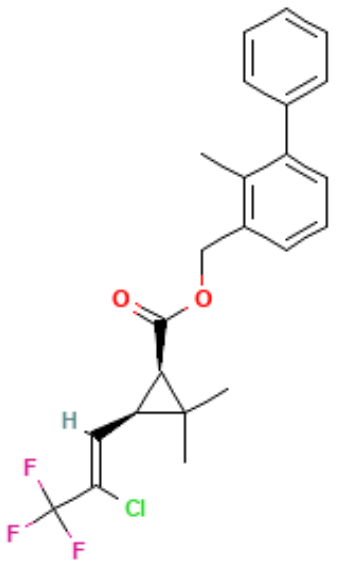

Рис. 4.7. Молекулярна структура Біфентрину

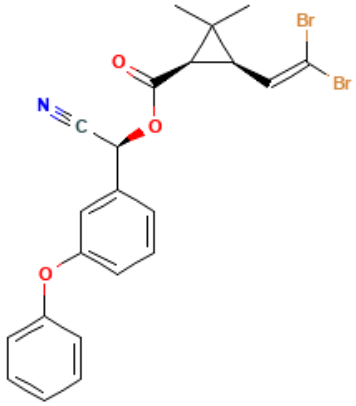

Рис 4.8. молекулярна структура Дельтаметрину

Взявши мас-спектр невстановленої молекули і прокласифікувавши його класифікатором ми можемо з певним ступенем ймовірності стверджувати, що ця молекула обов'язково містить ті чи інші структурні елементи. В даному випадку це будуть карбоксильна група рис. 4.9. а) і фенольне кільце рис. 4.9. б).

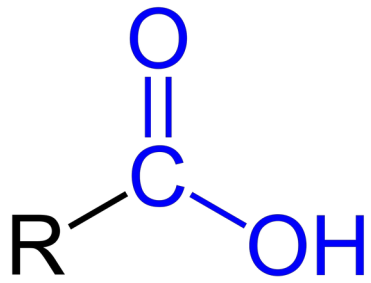

Рис. 4.9. Карбоксильна група

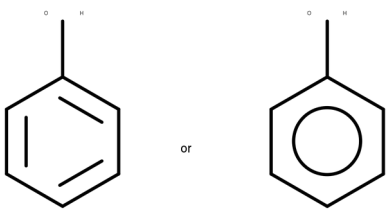

Рис. 4.10. Фенольне кільце

ВИСНОВКИ

Під час проведення даної роботи були розглянуті та проаналізовані найпопулярніші методи машинної класифікації. З метою вирішення нашої конкретної задачі - ідентифікації хлорорганічних молекул за допомогою класифікації їх мас-спектрів, було обрано машинний алгоритм - дерево рішень.

Результати були поділені на два типи кластеризації. Спочатку дані, інтенсивності піків мас-спектрів було кластеризовано вручну за допомогою вектора ознак, кінцевий розрахунок був відображений на гіперплощині, де було виявлено 10 класів. Для більш швидкого процесу, було використано інший підхід: в програмному середовищі R-Studio з використанням мови R було розроблено та реалізовано класифікатор за допомогою машинного алгоритму дерево рішень. Цей класифікатор досягає нашої мети, що полягає в ідентифікації невідомих хлорорганічних молекул та визначенні їхніх мас-спектрів приналежності до таких класів, як біфініл, імідазол, неонікотиноїд, піретроїд, поліхлорбіфеніл, сульфонілсечовина, триазол, фенілкарбамат, хлорацетамід і ХОП. Крім того, за допомогою кластеризації k-means ми зобразили наочно розподіл відомих молекул і прогнозували, як будуть класифіковані невідомі молекули.

Ці результати дослідження демонструють успішне використання методу дерева рішень та кластеризації k-means для ідентифікації та класифікації хлорорганічних молекул на основі їх мас-спектрів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? Alexandre Varnek and Igor Baskin. J. Chem. Inf. Model. 2012, 52, 6, 1413–1437

2. S.M. Ali, et al. Butitaxel analogues: synthesis and structure-activity relationships J. Med. Chem., 40 (1997), pp. 236-241

3. Cherkasov, et al. QSAR modeling: where have you been? Where are you going to? J. Med. Chem., 57 (2014), pp. 4977-5010

4. H. Kubinyi Free Wilson analysis. Theory, applications and its relationship to Hansch analysis Quant. Struct. Act. Relat., 7 (1988), pp. 121-133

5. R.P. Sheridan, S.K. Kearsley Why do we need so many chemical similarity search methods? Drug Discov. Today, 7 (2002), pp. 903-911

6. A.G. Maldonado, et al. Molecular similarity and diversity in chemoinformatics: from theory to applications Mol. Divers., 10 (2006), pp. 39-79

7. J. Bajorath Molecular similarity concepts for informatics applications Methods Mol. Biol., 1526 (2017), pp. 231-245

8. N.M. Nasrabadi Pattern recognition and machine learning J. Electron. Imag., 16 (2007), p. 049901

9. E. Kondratovich, et al. Transductive support vector machines: promising approach to model small and unbalanced datasets Mol. Inf., 32 (2013), pp. 261-266

10. Hyvarinen, E. Oja Independent component analysis: algorithms and applications Neural Netw., 13 (2000), pp. 411-430

11. Chuprina, et al. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers J. Chem. Inf. Model., 50 (2010), pp. 470-479

12. J.D. MacCuish, N.E. MacCuish Chemoinformatics applications of cluster analysis Comput. Mol. Sci., 4 (2014), pp. 34-48

13. L.B. Akella, D. DeCaprio Cheminformatics approaches to analyze diversity in compound screening libraries Curr. Opin. Chem. Biol., 14 (2010), pp. 325-330

14. Mohimani H., Gurevich A. et al. Dereplication of Peptidic Natural Products Through Database Search of Mass Spectra // Nature Chemical Biology. — 2016. — Vol. 13, no. 1. — P. 30–37.

15. Бейнон Дж. Мас-спектрометрія і її застосування в органічній хімії. Пер. з англ.. – М. Хімія, 1964.

16. Czamanski S. and Ablas L. A. Identification of industrial clusters and complexes: a comparison of methods and findings // Urban Studies. –1979. – V. I6 . – P. 61-80

17. Хмельницький Р. А., Мас-спектрометрія в органічній хімії. – Л.: Хімія, 1972.

18. Котов А., Красильников Н. Кластеризація даних. 2006.

19. Pattern Recognition and Machine Learning Крістофер Бішоп. Springer; 1st ed. 2006

20. http://www.cs.cmu.edu/~dpelleg/kmeans.html

21. W.S. Noble What is a support vector machine? Nat. Biotechnol., 24 (2006), pp. 1565-1567

22. A.C. Schierz Virtual screening of bioassay data J. Cheminf., 1 (2009), p. 21

23. F. Sahigara, et al. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions J. Cheminformatics, 5 (2013), p. 27

24. Svetnik, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling J. Chem. Inf. Comput. Sci., 43 (2003), pp. 1947-1958

25. https://de.wikipedia.org/wiki/RStudio

26. https://de.wikipedia.org/wiki/R_(Programmiersprache)

27. https://data-flair.training/blogs/clustering-in-r-tutorial/

ДОДАТОК 1

```
#####################################################################
############
################################ Класифікація
###################################
#####################################################################
############

# Робоча директорія
setwd("C:/Users/Робочий стіл/")

# Підключення даних
list.files()
library(readxl)
data <- read_excel("data.xlsx")
View(data)

# Подивимся датасет
dim(data)
sum(is.na(data))
data[1:20,456]
data[1:20,457]

# Видалим пропущенні значення
data <- na.omit(data)

# Видалим переміну name для дерева рішень
new_data <- data[,-457]

# Розіб'єбо вибірку на тренивочну та тестову
new_data_train <- new_data[1:1000,]
new_data_test <- new_data[-(1:1000),]

# Будуємо дерево рішень по тренивочній вибірці
library(rpart)
library(rpart.plot)
fit <- rpart(Clases ~., data = new_data_train)
prp(fit)
```

```
# Зроби прогноз на тестову вибірку
predict <- predict(fit, newdata = new_data_test, type = "class")

# Перевіримо точність моделі
correct_predictions <- sum(predict == new_data_test$Clases)
accuracy <- (correct_predictions / nrow(new_data_test)) * 100
accuracy

library(caret)
library(ggplot2)
library(ggthemes)
# Важливість переміних для дерева рішень
importance <- varImp(fit)

# Візуалізації важливості
imp_df <- data.frame(var = rownames(varImp(fit)),
                     importance = varImp(fit)$Overall)
plotlY <- ggplot(imp_df, aes(x = var, y = importance, fill = var)) +
  geom_col(show.legend = FALSE) +
  scale_y_continuous(expand = c(0,0)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5),
        plot.title = element_text(hjust = 0.5, size = 16),
        axis.title.x = element_text(size = 10)) +
  labs(title = "Графік важливості переміних",
       x = "Переміна",
       y = "Важливість")
plotlY

#####################################################################
############
################################ Кластеризація
###################################
#####################################################################
############
set.seed(123)
# Графік локтя
wss <- sapply(1:10, function(k) sum(kmeans(data[,1:455], centers=k)$withinss))
df <- data.frame(k=1:10, WSS=wss)
ggplot(df, aes(x=k, y=WSS)) +
  geom_point() +
```

```
  geom_line() +
  labs(title = "Графік локтя",
       x = "Кількість кластерів", y = "WSS") +
  theme_bw()
# Кластеризація kmeans
colors <- c("red", "blue", "green", "orange", "purple", "yellow", "cyan", "magenta",
"brown", "gray")
fit_clust <- kmeans(data[, 1:455], 10, iter.max = 200)
dist.data <- dist(data[, 1:455])
mds.data <- cmdscale(dist.data)
plot(mds.data, col = colors[fit_clust$cluster], pch = 20)
```

ДОДАТОК 2

X109 33.33333
X111 33.33333
X117 15.16712
X122 128.98453
X123 36.79244
X127 21.41176
X145 33.33333
X146 66.47121
X150 186.74996
X151 213.40640
X153 42.67064
X154 42.67064
X162 66.47121
X163 49.14910
X165 21.41176
X173 35.00000
X175 35.00000
X180 61.91042
X181 144.45485
X182 59.60940
X186 71.58706
X188 158.12364
X191 35.00000
X199 52.68110
X227 21.42626
X228 22.75036
X235 125.54722
X237 114.02037
X239 118.36210
X248 158.15425
X28   35.00000
X30 203.17606
X31   15.16712
X41 123.42680

X42 124.99667

X56 99.58114

X59 15.16712

X68 206.15983

X69 236.76354

X74 15.84727

X76 43.12132

X77 72.09222

X83 33.33333

X85 33.33333

X90 73.51841

X91 39.46348

X95 36.79244