МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ "ЛЬВІВСЬКА ПОЛІТЕХНІКА"

КУРСОВА РОБОТА

з дисципліни

“МАШИННЕ НАВЧАННЯ”

на тему: Моделювання поширення туберкульозу на території областей України на 2024 рік з визначенням найуразливіших груп за ознаками статі та віку

Студента 317 групи спеціальності

122 “Комп'ютерні науки”

Работягов Дмитро Сергійович

Керівник

к. е. н., доц. Бойко Н.І.

Кількість балів: Оцінка

Члени комісії

(підпис) (вчене звання, науковий ступінь, прізвище та ініціали)

(підпис) (вчене звання, науковий ступінь, прізвище та ініціали)

(підпис) (вчене звання, науковий ступінь, прізвище та ініціали)

Львів – 2024

ЗМІСТ

ВСТУП

Курсова робота присвячена актуальній проблемі моделювання поширення туберкульозу на території областей України у 2024 році з подальшим визначенням найуразливіших груп населення за ознаками статі та віку. Туберкульоз залишається однією із значущих глобальних проблем сучасного суспільства, особливо в умовах пандемії, і вимагає комплексного підходу до вивчення та контролю.

Наукова розробка цієї проблеми включає в себе різноманітні методи та підходи. На даний момент, вже використовуються статистичні методи, епідеміологічні моделі та методи машинного навчання для аналізу та прогнозування поширення хвороби. Проте, існують значні прогалини в розумінні динаміки та особливостей поширення туберкульозу на регіональному рівні в Україні.

Метою цієї роботи є вирішення цих проблем шляхом застосування сучасних методів машинного навчання та аналізу даних для побудови прогностичних моделей поширення туберкульозу на регіональному рівні. Висновки, отримані в результаті цього дослідження, можуть бути важливим внеском у розробку ефективних стратегій контролю та профілактики хвороби в Україні.

Завданням дослідження є:

1. Зібрати та підготувати дані: Оцінити наявні дані про кількість захворювань туберкульозом на території кожної області України за 2024 рік, враховуючи розподіл за віком та статтю.
2. Провести аналіз даних: Вивчити розподіл кількості захворювань за областями, віком та статтю для виявлення можливих залежностей та взаємозв'язків.
3. Розробити прогностичну модель: Застосувати методи машинного навчання, зокрема алгоритм Random Forest, для побудови прогностичної моделі поширення туберкульозу на регіональному рівні. Врахувати фактори ризику, такі як вік та стать, для визначення найуразливіших груп населення.

4. Оцінити ефективність моделі: Провести аналіз результатів та оцінити точність та надійність прогностичної моделі. Виявити найбільш вразливі групи населення.
5. Підготувати звітну документацію: Підготувати звіт з результатами дослідження та висновками, які можуть стати важливим внеском у розробку стратегій контролю та профілактики туберкульозу в Україні.

Переваги моделювання для такого типу задач полягають у здатності прогнозувати динаміку поширення хвороби, ідентифікувати найбільш ризиковані групи населення та виявлення ефективних стратегій контролю. Моделі дозволяють врахувати різноманітні фактори, такі як демографічні та соціально-економічні характеристики, що сприяє розумінню складних зв'язків, впливаючих на поширення хвороби. Також вони можуть бути корисним інструментом для прийняття рішень та розробки стратегій з превентивних заходів та лікування.

Актуальність теми відображає важливість і суттєвість проблеми, яку ми досліджуємо, та її відповідність сучасним потребам науки та практики. Моделювання поширення туберкульозу на території України є особливо актуальним у зв'язку зі зростанням числа випадків захворювання, зокрема у 2023 році. Дослідження спричинене високим рівнем захворюваності та потенційною загрозою громадського здоров'я є важливою проблемою, яка потребує ретельного аналізу та ефективних стратегій управління.

Розуміння поширення туберкульозу та виявлення найбільш уразливих груп населення є важливим для подальшого контролю та профілактики цієї хвороби. Дослідження в цій галузі не лише забезпечує наукову складову, але й має прямий практичний вплив на здоров'я громадян та систему охорони здоров'я країни.

Крім того, моделювання поширення туберкульозу з визначенням найуразливіших груп населення за ознаками статі та віку має значення для подальшого розвитку медичної науки та практики. Відкриття нових зв'язків та факторів, що впливають на розповсюдження хвороби, що може сприяти вдосконаленню методів діагностики та лікування туберкульозу, а також розробці

ефективних програм контролю та профілактики.

Об'єктом дослідження є процеси соціально-демографічних аспектів поширення туберкульозу за ознаками статі та віку. Предметом дослідження. є методи та алгоритми дослідження поширення туберкульозу на території областей України.

Основна увага дослідження спрямована на виявлення зв'язків між різними соціальними та демографічними факторами та розповсюдженням хвороби, а також на визначення чинників, що сприяють ризику захворювання серед різних груп населення. Таким чином, об'єкт і предмет дослідження відображають ключові аспекти, які досліджуються у рамках даної роботи з метою розв'язання проблеми поширення туберкульозу та забезпечення покращення громадського здоров'я.

Висновки до розділу

У розділі “Вступ” курсової роботи був проведений аналіз поширення хвороби туберкульозу, для того, аби зрозуміти реальну необхідність проведення операції моделювання поширення цього захворювання. Також, було чітко окреслено мету проведення дослідження, яка полягає в тому, щоб провести оцінку кількості нових захворювань серед населення кожної з областей України, та виділити суспільну групу людей, за категоріями віку та статі, які будуть найбільше уражені хворобою. Крім того, в даному розділі був проведений аналіз на актуальність обраної проблеми в суспільстві та виділення об’єкту та предмету дослідження.

Розділ 1. Аналіз літературних джерел

Для проведення дослідження використовуватимуться різноманітні методи, спрямовані на аналіз соціально-демографічних аспектів поширення туберкульозу та визначення найуразливіших груп населення. Один з основних методів - аналіз статистичних даних, який дозволить отримати об'єктивні результати щодо розподілу захворювання серед різних категорій населення. Також буде використано методи машинного навчання, такі як алгоритми регресії та кластеризація, для ідентифікації складних зв'язків та патернів у даних. Для аналізу географічного розподілу захворювання будуть використані набори даних про різні регіони за тривалий проміжок часу. Крім того, планується застосування епідеміологічних моделей для прогнозування поширення туберкульозу та оцінки ефективності стратегій контролю. Такий комплексний підхід до дослідження дозволить отримати глибше розуміння проблеми та визначити оптимальні шляхи боротьби з цією хворобою.

Ільницький Г. І. (2021) [https://lpnu.ua/sites/default/files/2021/radaphd/9119/ilnickiy-gi-dis.pdf] у своїй дисертації дослідив епідеміологічну ситуацію з туберкульозом в Україні. Автор, окрім аналізу статистичних даних та проведення соціологічних досліджень, також розробив математичну коміркову модель SIS для прогнозування поширення туберкульозу в Україні. Методи ґрунтуються на системі диференціальних рівнянь, які описують динаміку поширення туберкульозу в популяції. Методи враховують такі фактори, як: рівень захворюваності на туберкульоз, смертність від туберкульозу, рівень інфікування та ефективність лікування. Такий підхід не є дуже підходящим для умов реального світу, адже передбачає, що індивід стає інфекційним одразу після зараження, що не відповідає реальності, адже така хвороба має інкубаційний період. Також автор розглянув модель SEIS. Модель SEIS, яка розглядається тут, може бути інтерпретована як модель SIS із ефективною затримкою поширення захворювання. Вона показала себе краще, в умовах, наближених до умов реального світу.

Дослідником Роговським В. О. (2023)

[https://ela.kpi.ua/bitstream/123456789/29479/1/Rohovskyi_bakalavr.docx] було розроблено математичну математичну модель для прогнозування успішності лікування туберкульозу. Модель ґрунтується на системі диференціальних рівнянь, які описують динаміку кількості збудників туберкульозу в організмі. Модель враховує такі фактори: чутливість збудника до хіміотерапевтичних препаратів, імунний статус пацієнта, дотримання пацієнтом режиму хіміотерапії. Як було проаналізовано автором, модель достатньо добре робить передбачення на короткі проміжки часу, про що свідчить допустима похибка, проте застосування на більш тривалих періодах часу може призвести до збільшення похибки передбачення.

За даними Центру громадського здоров'я МОЗ України [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiFy6WjrtiEAxVy9wIHHTkrBeEQFnoECA8QAQ&url=https%3A%2F%2Fphc.org.ua%2Fsites%2Fdefault%2Ffiles%2Fusers%2Fuser90%2FTB_surveillance_statistical-information_2021_dovidnyk.docx&usg=AOvVaw3IhYDYNB5iE3viqs0PPTV1&opi=89978449], у 2022 році було зареєстровано 23 788 нових випадків туберкульозу, що становить 60,1 на 100 тис. населення. Очікування щодо зниження захворюваності та смертності від туберкульозу на 2022 рік: зниження захворюваності на 10%, зниження смертності на 5%. Реальні результати: зниження захворюваності на 8%, зниження смертності на 7%.

У розділі "Створення Байєсівської мережі факторів ризику захворіти COVID-19" дипломної роботи Шевченко Ярослава [http://ekhsuir.kspu.edu/bitstream/handle/123456789/16815/Шевченко%20Ярослава_211М_Дипломна%20Робота.pdf?sequence=1&isAllowed=y]описується два методи, що використовуються для побудови Байєсівської мережі: метод експертних оцінок (автор опитує 10 експертів (лікарів-інфекціоністів) щодо їх думки про вплив 12 факторів ризику на ймовірність захворіти COVID-19). На основі цих думок будується Байєсовська мережа. Плюси: простота та доступність(метод експертних оцінок простий у застосуванні та не потребує спеціальних знань з програмування), гнучкість(Байєсівські мережі легко адаптуються до нових даних та інформації), візуальність(Байєсівські мережі надають зручне візуальне представлення

взаємозв'язків між факторами ризику). Мінуси: суб'єктивність(метод експертних оцінок може бути суб'єктивним, що залежить від думки експертів), необхідність даних(алгоритм навчання Байєсівської мережі потребує великого обсягу даних), складність інтерпретації(інтерпретація результатів Байєсівської мережі може бути складною для людей без спеціальних знань).

З журналу “Хімія, екологія та освіта”, розділу “Математичне моделювання поширення вірусних інфекцій в локальних урбоекосистемах” [https://reposit.nupp.edu.ua/bitstream/PoltNTU/7617/3/IV%20%20МНПК_71.pdf] можна дізнатися, пропроцес розповсюдження респіраторно-вірусних захворювань, що найчастіше передаються повітряно-крапельним шляхом. Зараження цим шляхом найбільш ймовірні в місцях густого скупчення людей. Важливо зауважити, що заражати інших людина починає раніше, ніж починає сама хворіти. Були використані такі методики моделювання: математична модель (автори розробили математичну модель для прогнозування поширення респіраторно-вірусних інфекцій), динаміка Ланжевена(модель використовує динаміку Ланжевена для моделювання руху агентів в системі), моделювання польотів Леві(модель дозволяє враховувати раптові переміщення інфікованих агентів на великі відстані). Плюси: адекватність(модель адекватно відображає основні просторово-часові складові урбоекосистеми), універсальність(модель універсальна і може бути налаштована відповідно до потреб), ефективність(модель може використовуватися для прогнозування епідемій та оцінки ефективності методів профілактики). Мінуси: складність(модель може бути складною для розуміння та використання), необхідність великого обсягу даних(для параметризації моделі потрібні дані про урбоекосистему та поведінку людей), обмеження(модель не може врахувати всі фактори, що впливають на поширення інфекцій).

Штучний інтелект може стати потужним інструментом для прогнозування поширення ТБ та розробки ефективних стратегій його контролю. Стаття під назвою “Перспективи застосування штучного інтелекту для прогнозування поширення туберкульозної інфекції в Європейському регіоні ВООЗ”

[http://tubvil.com.ua/article/view/282415/276650] тільки це підтверджує. В дослідженні автори описують різні методи ШІ, які можуть бути використані для прогнозування ТБ. Моделі поширення туберкульозу: класична SIR-модель:

a) Модель використовує три стани для агентів: сприйнятливий, інфікований, одужав.
b) Модель враховує такі фактори, як: кількість агентів, швидкість руху, ймовірність інфікування, тривалість хвороби тощо.

Модель міського середовища:

c) Автори пропонують розробити модель, яка враховує життя, поведінку та взаємодію людей в місті.
d) Ця модель буде інтегрована з моделлю поширення інфекції для більш точного прогнозування.

Плюси запропонованого підходу:

Кращий прогноз: Комбінація моделей дозволить більш точно прогнозувати поширення епідемії на рівні регіону, країни та світу.

Врахування геопросторових даних: Використання географічних карт та розташування будівель дозволить досліджувати просторове поширення епідемії.

Швидкість розрахунків: Модель повинна бути швидкодійною для проведення великої кількості комп'ютерних експериментів.

Паралельні обчислення: Можливість паралельних обчислень дозволить моделювати поширення епідемії на макрорівні держав та світу.

Мінуси запропонованого підходу:

Складність розробки: Розробка моделі міського середовища є складним завданням.

Модель повинна враховувати багато факторів, таких як вік, імунітет, типи будівель тощо.

Необхідність даних: Для розробки та навчання моделі потрібні великі обсяги даних.

У таблиці 1 подано аналіз досліджень з даної теми, який підлягає

ретельному науковому розгляду з метою розуміння напряму розвитку змін.

Таблиця 1

Розгляд суміжних робіт

Назва роботи(автор)	Методика	Плюси методики	Мінуси методики
Ільницький Г. І. (2021)	Математична коміркова модель SIS/SEIS	Проста, добре описує динаміку поширення туберкульозу	Не враховує інкубаційний період, не дуже підходить для реального світу
Роговський В. О. (2023)	Математична модель динаміки кількості збудників туберкульозу	Дозволяє враховувати чутливість до хіміотерапії, імунний статус, дотримання режиму	Не дуже точна на довгих проміжках часу
Центр громадського здоров'я МОЗ України (2022)	Очікування та реальні результати щодо зниження захворюваності та смертності від туберкульозу	Простий метод прогнозування	Не враховує вплив різних факторів
Шевченко Ярослава	Байєсовська мережа факторів ризику захворіти COVID-19	Проста, гнучка, візуальна	Суб'єктивна, потребує багато даних, складна для інтерпретації
Журнал “Хімія, екологія та освіта” (2023)	Математична модель, динаміка Ланжевена, моделювання польотів Леві	Адекватна, універсальна, ефективна	Складна, потребує багато даних, має обмеження
Стаття “Перспективи застосування	Класична SIR-модель, модель міського	Кращий прогноз, врахування геопросторових	Складність розробки, потреба в багатьох даних

штучного інтелекту для прогнозування поширення туберкульозної інфекції в Європейському регіоні ВООЗ”	середовища	даних, швидкість розрахунків, можливість паралельних обчислень	

Підсумовуючи результати наведені у табл. 1, можна зробити висновок з аналізу літературних джерел, що підтверджує актуальність проблеми вивчення туберкульозу в Україні. Незважаючи на певний спад захворюваності протягом останніх років, рівень туберкульозу в країні залишається вищим, ніж у країнах Європейського Союзу.

Математичні моделі виявляються корисним інструментом для прогнозування поширення захворювання та оцінки ефективності лікування. Проте, важливо розуміти, що ці моделі мають свої обмеження і не завжди точно відображають реальну ситуацію. Додаткові дослідження та вдосконалення методів аналізу можуть сприяти покращенню точності прогнозів та ефективності стратегій боротьби з туберкульозом в майбутньому.

Висновок до розділу

В цьому розділі необхідно було дослідити поточний стан вирішення обраної проблеми, для чого був проведений аналіз літературних джерел, де ми виділили роботи інших дослідників на схожі теми. Розглядаючи ці роботи, важливим завданням було виділити використані ними алгоритми та методи, після чого з’ясувати які переваги та недоліки виділили автори для використаних методів, з чого в подальшому була сформована таблиця 1.

Розділ 2. Аналіз матеріалів та методів

2.1. Опис набору даних

Для передбачення поширення туберкульозу та його моделювання, необхідним є набір даних, який буде містити достатню кількість інформації, аби передбачення мали високу точність. Саме тому, для дослідження було обрано набір [https://drive.google.com/drive/folders/1RpivIgIFGlvTCNRJXSwY3A0-FRySJHvg], який містить дані про захворюваність на туберкульоз в Україні за тривалий період з 2007-2022 роки. Дані представлені у табличному форматі і описують такі характеристики: кількість захворювань за параметрами дати та області, віку та статі, форма туберкульозу та результату лікування.

Хоча датасет не містить прямої інформації про причини виникнення туберкульозу, проте він дає можливість дослідити фактори ризику, такі як вік, місце проживання, соціально-економічний статус та наявність супутніх захворювань.

Дані описують як чоловіків, так і жінок всіх вікових груп, з детальною категоризацією віку.

Окрім інформації про форми туберкульозу (легенева, позалегенева), датасет також містить дані про результати лікування, що дозволяє оцінити їх ефективність.

Наявний датасет буде розділено у відношенні 7:3 для тренування та тестування відповідно.

Постановка завдання

Для кращого розуміння підходів та для розв'язання поставлених завдань розглянемо запропоновані методи, наведені в Табл. 1

Методика математичної коміркової моделі SIS/SEIS, описана Ільницьким Г. І. (2021), серед переваг має простоту використання та гарний опис динаміки поширення захворювання, проте, важливим негативним фактором методики є те, що вона не враховує інкубаційний період захворювання, через що її використання в даній роботі є недоречним.

Дослідником Роговським В. О. (2023) була запропонована математична

модель динаміки кількості збудників туберкульозу, яка чудово дозволяє враховувати чутливість до хіміотерапії, імунний статус, дотримання режиму, проте відсутність потреби у залученні цих параметрів до аналізу, разом з недостатньою точністю передбачень на довгих проміжках часу, роблять цю методику не підходящою для використання в роботі.

Дані, надані Центром громадського здоров'я МОЗ України (2022), лише показують порівняльну статистику по очікуваним і реальним результатом зниження захворюваності та смертності від туберкульозу.

В дипломній роботі Шевченко Ярослава, для вирішення поставленого питання, запропонована Байєсовська мережа факторів ризику захворіти COVID-19. Перевагами цього методу є гнучкість моделі, та гарна можливість візуалізувати отримані результати, проте на перевагу цьому, модель потребує набору даних значних розмірів, що протирічить зазначеному вище опису набору даних, та результати роботи алгоритму є складними для інтерпретації.

З проблемою необхідності набору даних великих розмірів, також зіштовхнулися методики динаміки Ланжевена та моделювання польотів Леві, що були описані в журналі “Хімія, екологія та освіта” (2023), хоча і мали такі переваги як універсальність роботи алгоритму та його ефективність.

В статті “Перспективи застосування штучного інтелекту для прогнозування поширення туберкульозної інфекції в Європейському регіоні ВООЗ” представлена класична SIR-модель, модель міського середовища, що має такі переваги як врахування геопросторових даних, швидкість розрахунків, можливість паралельних обчислень. Наявні також і недоліки методики, серед яких є складність розробки та потреба в наборі даних середньої - великої розмірності.

Проте для порівняльного аналізу слід розглянути описані в Табл. 1 методики, а також додаткові засоби, які найкращим чином підходять для вирішення поставленої задачі.

2.2. Дерево прийняття рішень(ДПР)

Для вирішення питання передбачення найуразливіших груп населення за категоріями віку та статі по захворюваності на туберкульоз, можна використати

дерева прийняття рішень, а саме Random Forest.Random Forest - це алгоритм машинного навчання, який використовує комбінацію ДПР для підвищення точності. Він добре працює з наборами даних середньої розмірності, адже саме в нашому випадку, вибраний набір даних містить інформацію по кожній з 24 областей України та міста Києва в часовому проміжку з 2007-2022 роки, що дає можливість прослідкувати динаміку поширення захворювання в кожній з областей протягом середнього за тривалістю проміжку часу в 15 років.

Random Forest може допомогти нам для:

1) Виявлення факторів ризику: Random Forest може допомогти визначити фактори (стать, вік, інші), які впливають на ймовірність захворювання на туберкульоз.
2) Класифікації людей на групи ризику: Random Forest може класифікувати людей на групи ризику за ймовірністю захворювання.
3) Виявлення найуразливіших груп за статтю та віком: Random Forest може допомогти вам визначити групи населення, які найбільш схильні до ризику захворювання на туберкульоз.

Модель буде робити передбачення найуразливіших груп, за параметрами кількості захворювання, вікових категорій, статі захворюваних та року. У результаті опрацювання даних, модель буде подавати на вихід вікову категорію і стать людей, які можуть бути найбільш уражені хворобою.

Модель працює наступним чином:

1) Випадкова вибірка: З даних генерується множина випадкових підмножин.
2) Навчання ДПР: Для кожної підмножини даних будується ДПР.
3) Агрегування: Прогнози з усіх ДПР об'єднуються для отримання остаточного прогнозу.

Далі представлений псевдокод роботи алгоритму:

```
        Algorithm Random Forest: pseudocode
1   To generate c classifiers:
2   for i = 1 to c do
3         Randomly sample the training data D with replacement to produce D_i;
4         Create a root node, N_i, containing D_i,
5         Call BuildTree(N_i)
6   end for
7   BuildTree(N):
8   if N contains instances of only one class then
9         return
10  else
11        Randomly select x% of the possible splitting features in N
12        Select the feature F with the highest information gain to split on
13        Create f child nodes of N, N_1,.... , N_f, where F has f possible values
    (F_1, ..., F_f)
14        for i = 1 to f do
15           Set the contents of N_i, to D_i, where D_i, is all instances in N that
    match F_i
16            Call BuildTree(N_i)
17        end for
18  end if
```

Алгоритм роботи Random Forest:

1) Необхідно встановити такі параметри:
   a) Кількість дерев
   b) Глибина дерев
2) Далі необхідно навчити модель, для цього:

   для кожного дерева:

1) Виберемо випадкову підмножину даних
2) Навчимо ДПР на підмножині даних

3) Наступний етап це зробити прогноз, для цього:

для кожного дерева:

1) Зробимо прогноз для нового екземпляра даних
2) Об'єднаємо прогнози з усіх дерев

Також, для кращого розуміння роботи алгоритму, можемо переглянути блок схему його роботи на рис. 2.1:

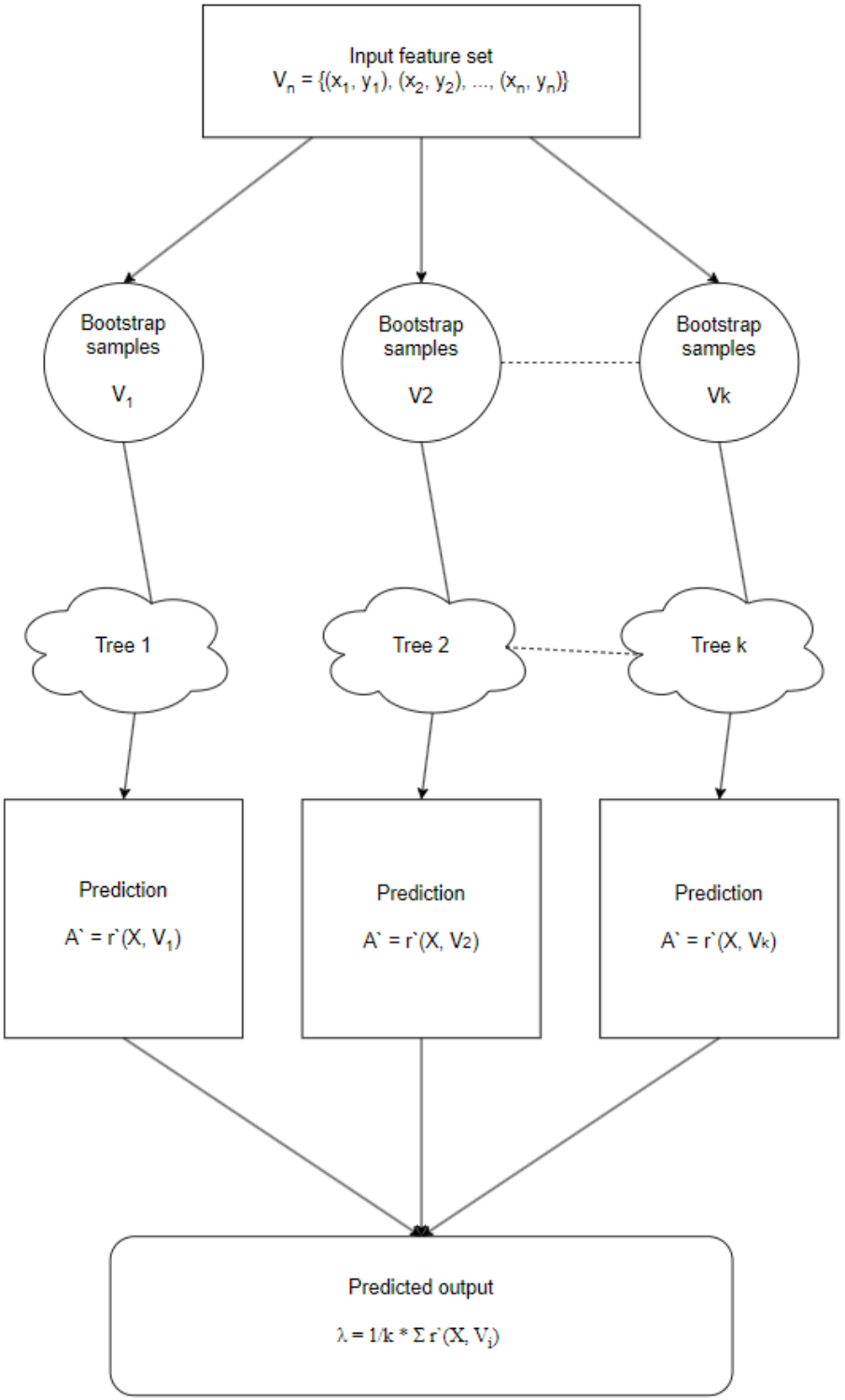

Рис. 2.1 Блок схема роботи алгоритму Random Forest

- На рис. 2.1 є такі умовні позначення:

  V_n: Це весь набір даних, який використовується для навчання алгоритму. Він складається з n точок даних, де кожна точка даних представлена парою (x_i, y_i). Тут x_i - це вектор ознак для i-ї точки даних, а y_i - це відповідне значення цілі.

- v: Це число дерев рішень, яке буде створено в алгоритмі.
- V_1, V_2, ..., V_k: Це випадкові набори даних, які використовуються для навчання кожного дерева рішень. Кожен набір даних V_i складається з k точок даних, які випадково вибрані з V_n.
- k: Це число ознак, які випадково вибираються для кожного дерева рішень.
- A_1, A_2, ..., A_k: Це прогнози, зроблені кожним деревом рішень для нового екземпляра даних X.
- A: Це остаточний прогноз, зроблений алгоритмом Random Forest. Він обчислюється як середнє прогнозів, зроблених всіма деревами рішень.

Для оцінки точності можна використати R^2-міру(коефіцієнт детермінації), а для оцінки похибки моделі середньоквадратичну похибка(середнє значення квадратів різниць між прогнозами та дійсними значеннями), що далі позначатиметься як MSE.

2.3. Модель клітинних автоматів

Наступним етапом після визначення найуразливіших груп населення виступає моделювання поширення туберкульозу на території областей України. Для виконання цієї частини завдання було вирішено зупинитися на Моделі SIR. Модель SIR - це проста модель клітинних автоматів, яка використовується для моделювання поширення інфекції. Вона розділяє людей на три групи:

1) Сприйнятливі (S): Люди, які можуть заразитися інфекцією.
2) Інфіковані (I): Люди, які заражені інфекцією і можуть передавати її іншим.
3) Ті хто одужали (R): Люди, які одужали від інфекції і більше не можуть заразитися.

Модель SIR працює наступним чином:

Інфіковані люди передають інфекцію сприйнятливим людям з певною ймовірністю.

Сприйнятливі люди, які заражаються інфекцією, стають інфікованими.

Інфіковані люди з часом одужують.

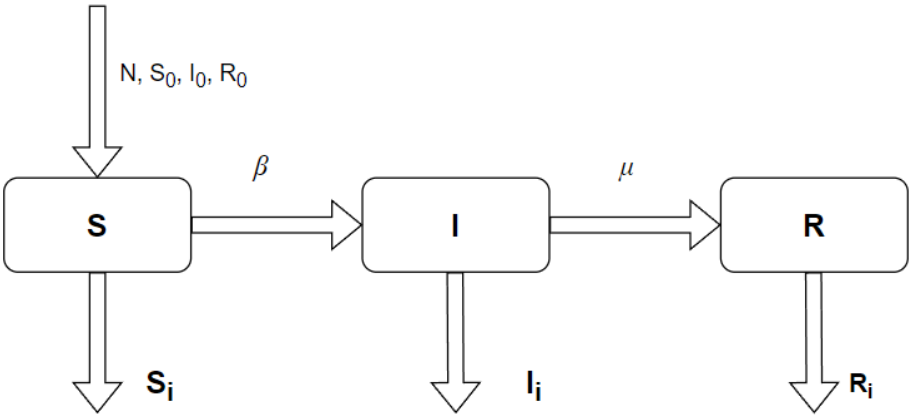

Рис. 2.2 Компартментарна діаграма моделі SIR

На рис. 2.2 наведено зображення компраметарної діаграми моделі SIR, де N, S_0, I_0, R_0 це загальна кількість населення, початкове значення кількості людей, що можуть захворіти, початкове значення заражених людей та початкове значення людей що більше не можуть захворіти відповідно, β - це символ, що позначає коефіцієнт інфікування(описує ймовірність того, що сприйнятливий індивідуум заразиться при контакті з інфікованим індивідом), μ - це символ, що позначає коефіцієнт смертності/одужання(описує ймовірність того, що інфіковані індивідууми помруть від захворювання або видужають без можливості знову захворіти).

Значення S_i, I_i та R_i обраховуються на основі диференційних рівнянь на кожній ітерації моделювання наступним чином відповідно:

$$S_i = \frac{-\beta}{N * S_{i-1} * I_{i-1}}$$

$$I_i = \frac{\beta}{N * S_{i-1} * I_{i-1} * (I - \mu)}$$

$$R_i = I * \mu$$

Нижче представлений псевдокод роботи алгоритму:

Algorithm SIR: pseudocode

```
1   function MAIN
2       set t, fixed = (p^(fix), t), bounds = range for p^(cal)
3       for run = 0 to Number_of runs parallel do
4           call single run (bounds, fixed)
5       end for
6       read and assembles results from single runs
7   end function
8   function SINGLE_RUN(bounds, fixed)
9       call differential_evolution (Objective function, bounds,
    fixed)
10      save results of a single run
11  end function
12  function OBJECTIVE_FUNCTION(p, t)
13      s = SIRmodel (p)
14      objfun=O_y(s),t(p)
15      return objfun
16  end function
17  function SIRMODEL(p)
18      initialize S_nini
19      for n = n_ini + 1 to n_ini + N^(mod) - 1 do
20          compute s_n from s_n-1 by (10)
21      end for
22      return s
23      end function
```

2.4. Підготовка даних

Перед тим як ділити дані на тренувальні і тестувальні для подальшої обробки алгоритмами, необхідно привести дані до одного стандарту. Для цього, необхідно викинути таблиці, які не будуть використовуватися при аналізі, а саме Блок 4. Штати протитуберкульозних закладів, Блок 5. Лікування туберкульозу, Блок 6. Ліжковий фонд протитуберкульозних закладів, та інші окремі таблиці. Інші таблиці, необхідно поєднати за ознаками років та областей, при тому очистивши дані(заповнення пропусків за наявності, нормалізація, стандартизація і тд). Приклади блоків таблиць, розділених за інформацією, яка в них зберігається наведені на рис. 2.3:

	Блок 3. Туберкульоз/ВІЛ-інфекція.
Таблиця 44	Захворюваність на туберкульоз у поєднанні зі СНІДом (нові випадки+рецидиви)
Таблиця 45	Реєстрація ВІЛ-позитивних осіб, хворих на туберкульоз
Таблиця 46	Померло хворих на туберкульоз від хвороби, зумовленої СНІДом
Таблиця 47	Поширеність всіх форм активного туберкульозу у поєднанні з хворобою, зумовленою ВІЛ
	Блок 4. Штати протитуберкульозних закладів.
Таблиця 48	Забезпеченість лікарями-фтизіатрами у закладах системи МОЗ України
Таблиця 49	Медичні посади у лікувально-профілактичних закладах системи МОЗ України, 2021 рік

	Блок 6. Ліжковий фонд протитуберкульозних закладів.
Таблиця 69	Мережа протитуберкульозних закладів охорони здоров'я системи МОЗ України та кількість ліжок для хворих на туберкульоз, 2021 рік
Таблиця 70	Забезпеченість лікарняними ліжками для хворих на туберкульоз у закладах охорони здоров'я системи МОЗ України
Таблиця 71	Показники використання ліжкового фонду протитуберкульозних закладів охорони здоров'я системи МОЗ України, 2021 рік
Таблиця 72	Лікарняна та санаторна допомога хворим на туберкульоз відповідно до територіального розміщення закладів охорони здоров'я системи МОЗ України

Рис. 2.3 Приклади блоків таблиць з вибраного набору даних

Для виконання роботи необхідно спроектувати 2 набори даних. Перший набір, що буде використовуватися для передбачення вікової та статевої категорій людей, що найбільш вразливі для подальшого зараження, за допомогою алгоритму Random Forest, буде складатися з таких колонок: AgeCategory, Year, Sex, TotalPopulation, Infected, Dead. Приклад з цього набору даних зображений на рис. 2.4:

AgeCategory	Sex	Year	TotalPopulation	Infected	Dead
0-1	M	2007	46 650 000	15	2
1-4	M	2007	46 650 000	113	0
5-9	M	2007	46 650 000	107	1
10-14	M	2007	46 650 000	99	1
15-17	M	2007	46 650 000	300	1
18-24	M	2007	46 650 000	2445	162
25-34	M	2007	46 650 000	6195	1365
35-44	M	2007	46 650 000	6310	2405
45-54	M	2007	46 650 000	5958	2890
55-64	M	2007	46 650 000	2831	1195
65 - 100	M	2007	46 650 000	1895	586
0-1	F	2007	46 650 000	7	1
1-4	F	2007	46 650 000	83	3
5-9	F	2007	46 650 000	90	0
10-14	F	2007	46 650 000	107	1
15-17	F	2007	46 650 000	303	1
18-24	F	2007	46 650 000	1707	91
25-34	F	2007	46 650 000	2821	464
35-44	F	2007	46 650 000	2118	512
45-54	F	2007	46 650 000	1482	449
55-64	F	2007	46 650 000	806	171
65 - 100	F	2007	46 650 000	1303	205

Рис. 2.4. Приклад набору даних для алгоритму Random Forest

Опис елементів набору даних для алгоритму Random Forest

Таблиця 2.1

Назва колонки	Опис
AgeCategory	Вікові категорії, представлені в початковому наборі даних від 0 до 100 років
Sex	Стать
Year	Рік, для якого описані кількість інфікованих та померлих
TotalPopulation	Популяція України у вибраний рік
Infected	Кількість інфікованих людей вибраної статі та вікової категорії
Dead	Кількість померлих людей вибраної статі та вікової категорії

В таблиці 2.1 представлений опис ознак, що складають набір даних для алгоритму Random Forest.

Для набору, що буде використовуватися алгоритмом SIR для моделювання поширення захворюваності, спроектуємо датасет, що буде складатися з наступних колонок: Year, InfectedTB, DeadTB, Region, TotalPopulation, RecoveredTB. Приклад сформованих даних наведено на рис. 2.5:

Year	InfectedTB	DeadTB	Region	TotalPopulation	RecoveredTB
2022	405	93	Volyn	1 018 628	150
2022	2 450	340	Dnipropetrovsk	3 093 176	1013
2022	133	13	Donetsk	1 883 713	58
2022	491	98	Zhytomyr	1 179 801	189
2022	880	136	Zakarpattia	1 241 643	358
2022	595	161	Zaporizhzhia	1 637 673	209
2022	378	44	Ivano-Frankivsk	1 349 096	161
2022	857	129	Kyiv	1 789 300	350
2022	607	100	Kirovohrad	897 297	244
2022	13	78	Luhansk	666 801	-32
2022	846	195	Lviv	2 459 763	313
2022	952	85	Mykolaiv	1 091 106	417
2022	2 927	249	Odessa	2 340 332	1286
2022	700	70	Poltava	1 344 445	303
2022	530	67	Rivne	1 140 724	223
2022	271	88	Sumy	1 033 580	88
2022	280	29	Ternopil	1 018 462	121
2022	965	181	Kharkiv	2 583 325	377
2022	476	100	Kherson	1 000 166	181
2022	441	50	Khmelnytskyi	1 225 666	188
2022	419	95	Cherkasy	1 157 115	156
2022	406	42	Chernivtsi	887 392	175
2022	469	78	Chernihiv	950 773	188

Рис. 2.5 Приклад набору даних для моделі SIR

Опис елементів набору даних для алгоритму моделі SIR

Таблиця 2.2

Назва колонки	Опис
Year	Рік, для якого описані кількість інфікованих, реабілітованих та померлих
Region	Область України. Представлений у вигляді назви регіонального центру(крім окремо міста Київ)
RecoveredTB	Кількість одужавших людей вибраної області
TotalPopulation	Популяція України у вибраний рік

InfectedTB	Кількість інфікованих людей вибраної області
DeadTB	Кількість померлих людей вибраної області

В таблиці 2.2 наведений опис ознак, що входять до набору даних для алгоритму SIR.

Висновки до розділу

В процесі проведення аналізу матеріалів та методів, нами був виконаний аналіз наборів даних, які б найкраще підходили для вирішення поставленої задачі, серед яких був вибраний такий, який має інформацію по багатьох параметрах, зокрема по часовому проміжку з 2007 по 2022 роки. Після чого, необхідно було виділити переваги і недоліки серед методів, використаних іншими дослідниками, описаних в таблиці 1 розділу 1. Зважаючи на те, що серед представлених в таблиці методів не знайшлося підходящих для проведення моделювання поширеності захворювання та передбачення найвразливіших соціальних груп населення, з’явилася необхідність провести аналіз додаткових методів, таких як SIR модель клітинних автоматів та Random Forest, для вирішення поставлених задач відповідно. Для кожного з цих методів були описані алгоритми їх роботи, а також, описані початкові, та сформовані набори даних, що будуть використовуватися кожним з алгоритмів.

Розділ 3. Експерименти

Проведення експериментів для теми “Моделювання поширення туберкульозу на території областей України на 2024 рік з визначенням найуразливіших груп за ознаками статі та віку” має велике значення, адже дозволяє зрозуміти наскільки точно відпрацювали вибрані моделі чи алгоритми, а саме, зрозуміти наскільки точно модель RandomForest передбачає кількість інфікувань захворюванням в подальші роки, з чого можна зробити висновки про найуразливішу категорію населення. Крім того, експерименти дають можливість наглядно побачити змодельовану динаміку поширення захворюваності на кожній з областей країни за допомогою моделі клітинних автоматів SIR.

3.1 Засоби розробки

Програмну реалізацію було виконано за допомогою Python - це високорівнева мова програмування загального призначення, яка має простий та читабельний синтаксис. Вона має велику кількість бібліотек для різноманітних задач, що робить її дуже потужним інструментом для розв'язання проблем у наукових дослідженнях, аналізі даних та машинному навчанні. Переваги Python включають простоту вивчення, широку підтримку спільнотою та зручність у використанні. Зокрема, для розробки програми були використані наступні бібліотеки та інструменти:

- Pandas - це бібліотека для обробки та аналізу даних, яка надає структури даних та функції для роботи з ними. Вона дозволяє легко виконувати операції з великими наборами даних, такими як читання, запис, фільтрація та агрегація даних.
- NumPy - це бібліотека для наукових обчислень в Python. Вона надає підтримку для масивів та математичних функцій, що дозволяє легко виконувати розрахунки над числовими даними.
- Scikit-learn - це бібліотека машинного навчання для Python. Вона містить реалізації багатьох алгоритмів машинного навчання, таких як класифікація, регресія, кластеризація та інші, а також інструменти для оцінки та підбору

параметрів моделей.

- Seaborn - це бібліотека для візуалізації даних в Python, яка базується на бібліотеці Matplotlib. Вона надає високорівневі функції для створення привабливих та інформативних графіків та діаграм.
- Matplotlib - це бібліотека для створення графіків та візуалізації даних в Python. Вона дозволяє створювати різноманітні типи графіків, такі як лінійні, кругові та гістограми, та налаштовувати їх вигляд.
- Scipy - це бібліотека для наукових та технічних обчислень в Python. Вона містить реалізації багатьох алгоритмів для чисельного обчислення, оптимізації, обробки сигналів, а також інші функції для роботи з науковими даними.

3.2 Передбачення найуразлівішої групи населення за допомогою алгоритму Random Forest

Описаний в пункті 2.4 набір даних для алгоритму Random Forest, приклад якого можна бачити на рис. 2.4, зберігається у форматі `.csv,` тому необхідно написати функцію яка зчитає дані з файлу і поверне їх у форматі `pandas DataFrame:`

	Функція readDataFrame(*шлях до файлу*):
1	Ініціалізуємо порожній DataFrame під назвою ***df_rf***
2	Прочитаємо вміст файлу за допомогою «pd.read_csv()»
3	Додамо вміст прочитаного файлу до ***df_rf***
4	Повернемо ***df_rf***

	AgeCategory	Sex	Year	TotalPopulation	Infected	Dead
0	0-1	M	2007	46 650 000	15	2
1	1-4	M	2007	46 650 000	113	0
2	5-9	M	2007	46 650 000	107	1
3	10-14	M	2007	46 650 000	99	1
4	15-17	M	2007	46 650 000	300	1
...	...	...	...	...	...	...
347	25-34	F	2022	41 167 000	753	63
348	35-44	F	2022	41 167 000	1119	134
349	45-54	F	2022	41 167 000	952	121
350	55-64	F	2022	41 167 000	644	79
351	65 - 100	F	2022	41 167 000	787	77

Рис. 3.1 Набір даних у форматі pandas DataFrame

На рис. 3.1 представлений приклад даних із завантаженого набору даних. Після зчитання, необхідно перевірити типи даних, які зберігаються в наборі. Результат перевірки можна бачити на рис 3.2:

```
AgeCategory        object
Sex                object
Year                int64
TotalPopulation    object
Infected           object
Dead                int64
dtype: object
```

Рис. 3.2 Типи даних завантаженого набору даних

Для роботи моделі необхідно попередньо провести очищення даних, перевірку на пробіли, та переведення в необхідні типи, після чого набір даних має наступний вигляд

	AgeCategory	Sex	Year	TotalPopulation	Infected	Dead
0	0-1	M	2007	46650000	15	2
1	1-4	M	2007	46650000	113	0
2	5-9	M	2007	46650000	107	1
3	10-14	M	2007	46650000	99	1
4	15-17	M	2007	46650000	300	1
...	...	...	...	...	...	...
347	25-34	F	2022	41167000	753	63
348	35-44	F	2022	41167000	1119	134
349	45-54	F	2022	41167000	952	121
350	55-64	F	2022	41167000	644	79
351	65 - 100	F	2022	41167000	787	77

Рис. 3.3 Дані після очищення

На рис. 3.3 представлений приклад даних після проведення операцій по їх очищенню.

```
AgeCategory        object
Sex                object
Year                int64
TotalPopulation     int64
Infected            int64
Dead                int64
dtype: object
```

Рис. 3.4 Типи даних після очищення набору даних

На рис. 3.4 представлені типи даних, з набору даних для алгоритму Random Forest після проведення операцій очищення даних.

```
AgeCategory        0
Sex                0
Year               0
TotalPopulation    0
Infected           0
Dead               0
dtype: int64
```

Рис. 3.5 Перевірка на наявність пропусків в наборі даних

Як можна бачити з рис. 3.5, дані в наборі не мають пропусків, отже можна переходити до нормалізації необхідних колонок даних, а саме TotalPopulation, Infected та Dead, та проведення операції label encoding для колонок Sex та AgeCategory. Нормалізація буде відбуватися за допомогою наступної функції:

```
Функція min_max_normalize(колонка набору даних):
1    Шукаємо мінімальне значення в колонці і встановлюємо його в змінну col_min
2    Шукаємо максимальне значення в колонці і встановлюємо його в змінну col_max
3    Рахуємо оновленне значення елементів колонки за формулою (значення колонки - col_min) / (col_max - col_min)
4    Повертаємо колонку з оновленими значеннями
```

Операція label encoding буде відбуватися за допомогою функції `LabelEncoder`, яку ми імпортуємо з бібліотеки `sklearn.preprocessing`.

Після виконання зазначених вище операцій, новий вигляд набору даних для алгоритму Random Forest можна побачити на рис 3.6:

	AgeCategory	Sex	Year	TotalPopulation	Infected	Dead
0	0	1	2007	1.0	0.001458	0.000683
1	1	1	2007	1.0	0.015743	0.000000
2	8	1	2007	1.0	0.014869	0.000342
3	2	1	2007	1.0	0.013703	0.000342
4	3	1	2007	1.0	0.043003	0.000342
...	...	...	...	...	...	...
347	5	0	2022	0.0	0.109038	0.021524
348	6	0	2022	0.0	0.162391	0.045781
349	7	0	2022	0.0	0.138047	0.041339
350	9	0	2022	0.0	0.093149	0.026990
351	10	0	2022	0.0	0.113994	0.026307

Рис. 3.6 Вигляд набору даних після закінчення його підготовки

Для кращого розуміння залежності між даними в наборі даних, давайте проілюструємо таблицю кореляцій, зображену на рис. 3.7:

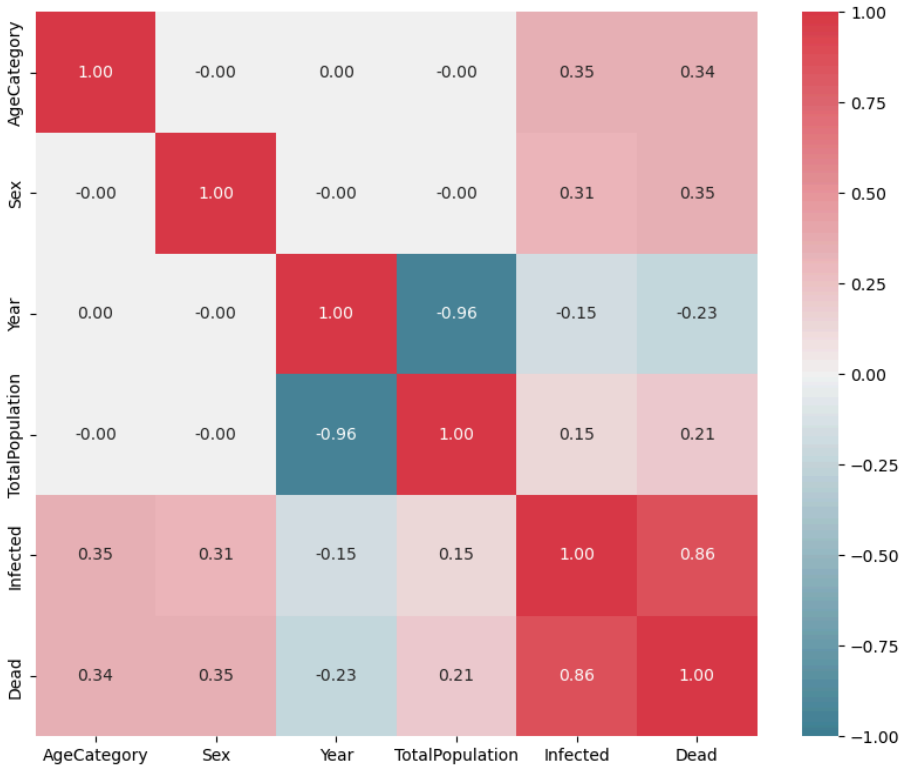

Рис. 3.7 Таблиця кореляцій набору даних

Сильною кореляцією будемо вважати клітинки з таблиці кореляції зі значеннями від 0.7 до 1.0, в той час як середньою буде вважатися кореляція, значення якої лежить в проміжку від 0.3 до 0.69. Як можемо бачити з рис. 3.7, ознаки Infected та Dead, ознаки AgeCategory та Infected, Sex та Infected мають середню кореляцію, в той час як ознаки TotalPopulation та Year мають сильну обернену кореляцію.

Перед початком роботи алгоритму, необхідно виділити цільову змінну, якою виступить ознака Infected, а всі інші ознаки будуть використовуватися для передбачення. Також, необхідно розділити набір даних на тренувальний а тестувальний, що буде зроблено у відношенні 7:3.

Нарешті, можемо переходити до моделі, яка буде використовуватися для передбачення. Зважаючи на те, що цільова ознака це неперервне значення, був використаний алгоритм `RandomForestRegressor`, взятий з бібліотеки мови Python `sklearn.ensemble`, де значення кількості дерев, що буде використане,

та передається як параметр встановлене як 100.

Після закінчення тренування моделі, запустимо тестування, та проведемо оцінку роботи моделі, за допомогою метрик середньоквадратичної помилки, та оцінки R^2. Отримані наступні результати:

```
Mean Squared Error (MSE): 0.0028440307441931815
R-squared: 0.9204798828368166
```

Як бачимо, значення середньоквадратичної помилки досить мале, в той час як значення оцінки R^2 достатньо високе, при найбільшому можливову значенні цієї оцінки 1.

Також, протестуємо роботу моделі, перевіривши яка група населення, за категоріями віку та статі, найбільш вразливі до зараження. Маємо наступні результати:

```
Based on the model, the group with the highest predicted
number of deaths in year 2024 is:
Age Category: 35-44
Sex: M
```

З цих результатів, маємо такий висновок, що чоловіки від 35 до 44 років, будуть мати найбільшу кількість заражень у подальший рік. Також, для кращого розуміння ситуації зараження для інших категорій статі та віку, виконаємо візуалізацію графіка передбачення поширення захворюваності серед них, що зображена на рис. 3.8:

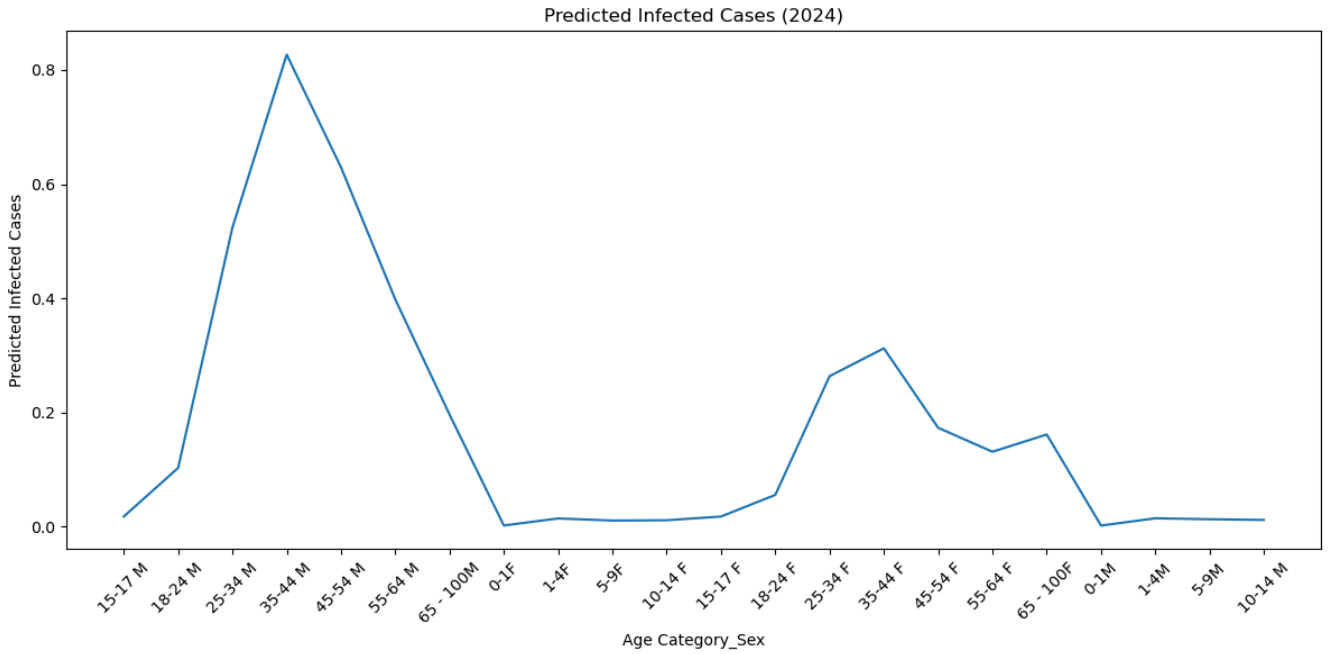

Рис. 3.8 Статистика подальших захворювань для всіх категорій статі та віку

З рис. 3.8 бачимо що вікова група чоловіків 35-44 років є найбільш вразивими до зараження, в той час як друге місце посідає група жінок тої ж вікової категорії. Значення для інших вікових категорій для обох статей також були проілюстровані на рис. 3.8.

3.3. Моделювання поширення туберкульозу на території областей України за допомогою моделі клітинних автоматів SIR

Для моделювання поширення захворювання на території окремо кожної з 24 областей України та окремо міста Київ, необхідно скористатися сформованим набором даних, описаним в пункті 2.4, приклад якого можна побачити на рис. 2.5. З прикладу цього набору даних, маємо інформацію про кількість заражених, кількість одужавших, кількість померлих, рік, назва області, та кількість населення відповідної області.

Завантажимо набір даних в програму, у форматі `pandas DataFrame` за допомогою функції `readDataFrame`, алгоритм роботи якої описаний в пункті 3.2.

	Year	InfectedTB	DeadTB	Region	TotalPopulation	RecoveredTB
411	2022	441	50	Khmelnytskyi	1 225 666	188
412	2022	419	95	Cherkasy	1 157 115	156
413	2022	406	42	Chernivtsi	887 392	175
414	2022	469	78	Chernihiv	950 773	188
415	2022	644	102	KyivCity	2 910 994	261

Рис. 3.9 Приклад набору даних у форматі `DataFrame`

На рис. 3.9 зображено приклад даних з набору даних для алгоритму SIR після переведення у формат `DataFrame`. Після зчитання, необхідно перевірити типи даних, які зберігаються в наборі, в результаті чого маємо такий результат:

```
Year                int64
InfectedTB         object
DeadTB              int64
Region             object
TotalPopulation    object
RecoveredTB         int64
dtype: object
```

Рис. 3.10 Типи даних завантаженого набору даних до очищення

З рис. 3.10 можемо бачити, що деякі числові ознаки, які мали б представляти числові значення не мають встановленого необхідного типу даних, для чого необхідно провести попередню очистку даних, та перевести необхідні ознаки у потрібні типи даних, після чого маємо:

```
Year                int64
InfectedTB          int64
DeadTB              int64
Region             string
TotalPopulation     int64
RecoveredTB         int64
dtype: object
```

Рис. 3.11 Типи даних завантаженого набору даних після очищення

На рис. 3.11 зображено опис типів даних для ознак набору даних для алгоритму SIR після проведення підготовчих операцій. Також, перевіримо дані на повноту, тобто чи не мають вони не заповнених полів. Результат перевірки

зображений на рис. 3.12:

```
Year              0
InfectedTB        0
DeadTB            0
Region            0
TotalPopulation   0
RecoveredTB       0
dtype: int64
```

Рис. 3.12 Перевірка набору даних на повноту

Як можемо бачити з рис. 3.12, дані не мають порожнин. Також, для кращого розуміння залежностей між ознаками в наборі даних, давайте проілюструємо таблицю кореляцій, зображену на рис. 3.13:

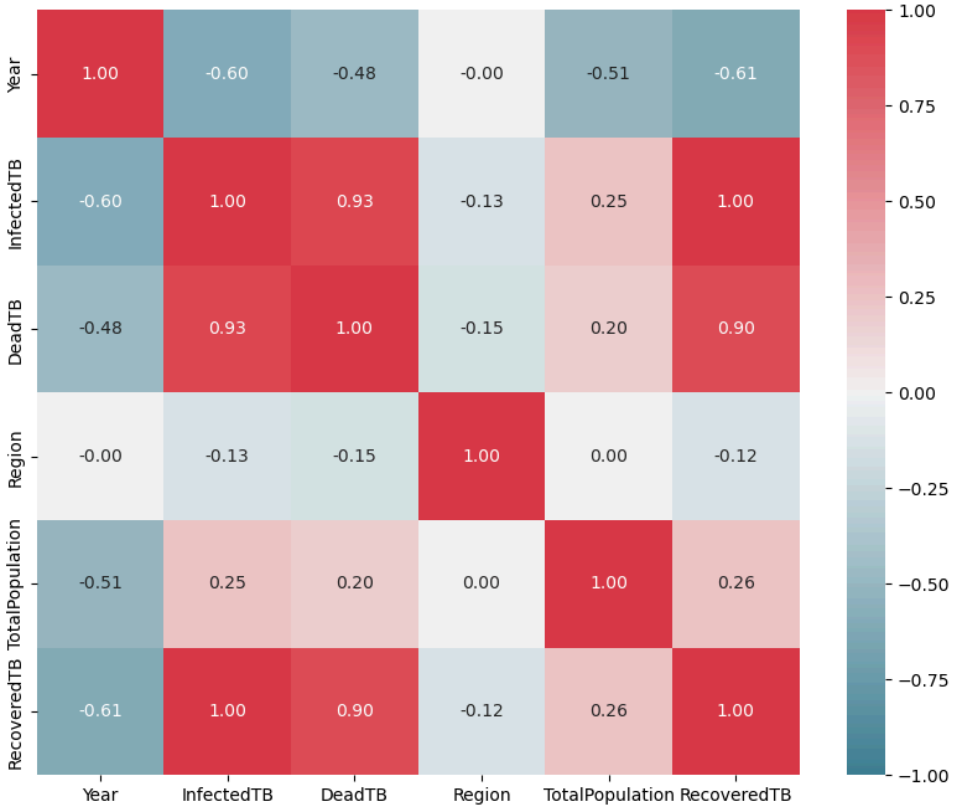

Рис. 3.13 Таблиця кореляцій для набору даних

В пункті 3.1, описуючи таблицю кореляцій для набору даних алгоритму Random Forest, зображену на рис. 3.7, ми надали визначення термінам сильна та середня кореляція, отже ознаки InfectedTB та RecoveredTB, DeadTB та RecoveredTB, InfectedTB та DeadTB мають сильну кореляцію, в той час як практично всі ознаки на перетині з ознакою Year мають сильну обернену

кореляцію.

Перед початком алгоритму SIR, необхідно описати початкові значення параметрів *S*, *I*, *R*, що означають відповідно кількість людей що можуть захворіти, кількість людей, що вже захворіли, та кількість людей, що більше не зможуть захворіти. В останню категорію я включив як людей що одужали, так і людей що вже померли від хвороби. Серед параметрів, також присутній один, який визначає кількість днів, протягом яких буде робитися моделювання поширення захворювання. В моєму випадку, він встановлений на 700 днів. Також, для роботи алгоритму, необхідно задати параметри β і μ які виступають коефіцієнтами при обрахунку кількості людей що переходять зі стану *S* в стан *I*(коефіцієнт передачі або швидкість, з якою схильні особи заражаються при контакті з інфікованими особами. Він вказує на швидшкість поширення хвороби в популяції), та зі стану *I* в стан *R*(швидкість одужання або темп, з яким інфіковані особи одужують від хвороби і стають імунними) відповідно. Отже, для вищеописаних параметрів, що необхідні для роботи алгоритму, маємо встановлені такі значення:

S = загальна кількість населення в області(TotalPopulation) - кількість інфікованих(InfectedTB) - кількість одужавших(RecoveredTB) - кількість померлих(DeadTB)

I = кількість інфікованих(InfectedTB)

R = кількість одужавших(RecoveredTB) + кількість померлих(DeadTB)

β = 4/10(що означає, що кожні 10 днів 4 особи заражаються)

μ = 1/10(що означає, що кожні 10 днів одужує одна особа)

Нарешті, після закінчення роботи алгоритму, маємо такі результати моделювання кількості придатних до захворювання, тих хто захворіли, та тих хто одужали або померли для деяких областей на рис. 3.14 - 3.16:

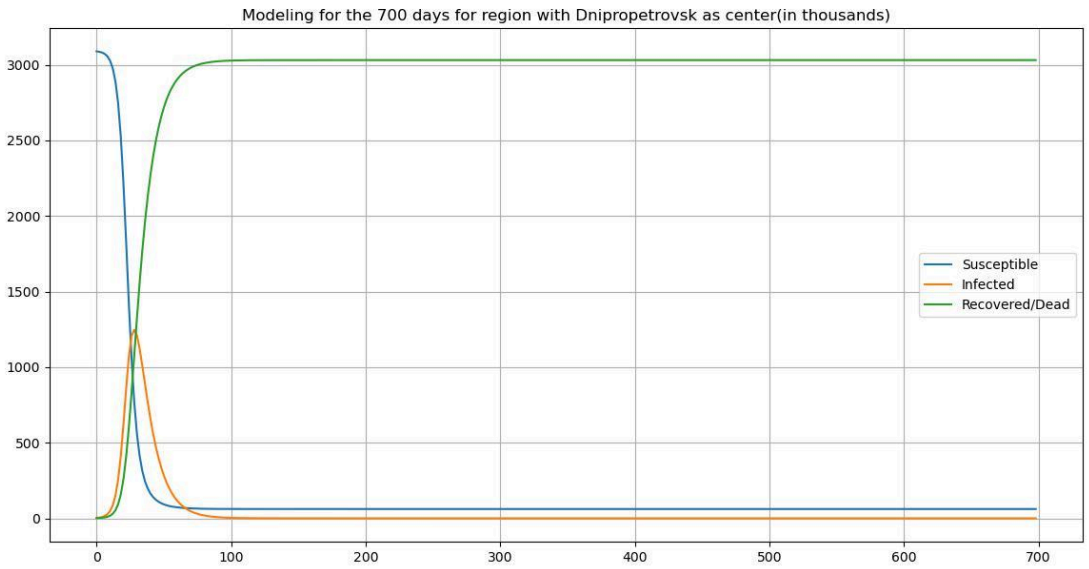

Рис 3.14 Результати моделювання для Дніпропетровської області

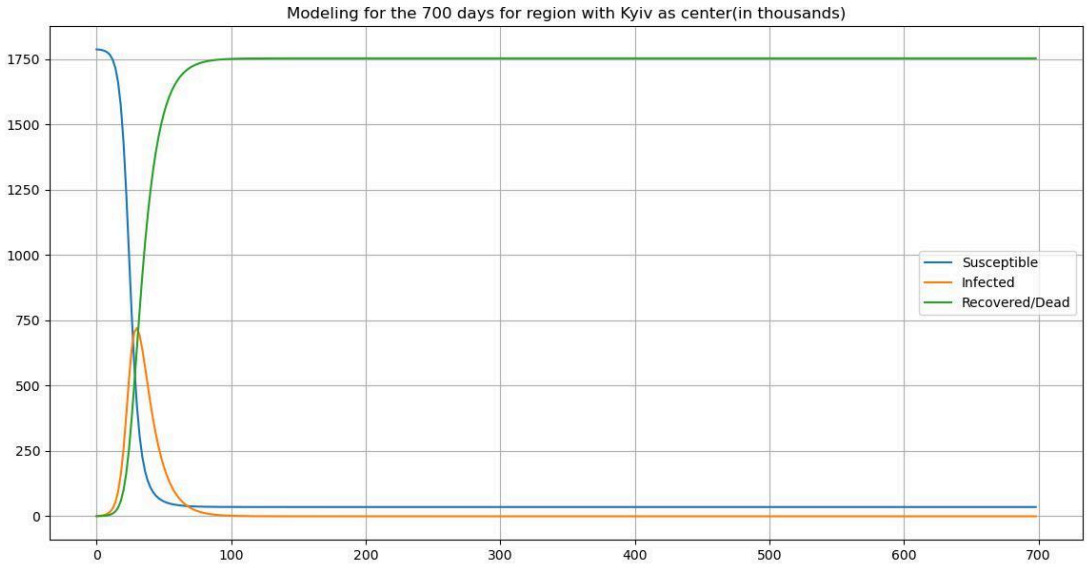

Рис 3.15 Результати моделювання для Київської області

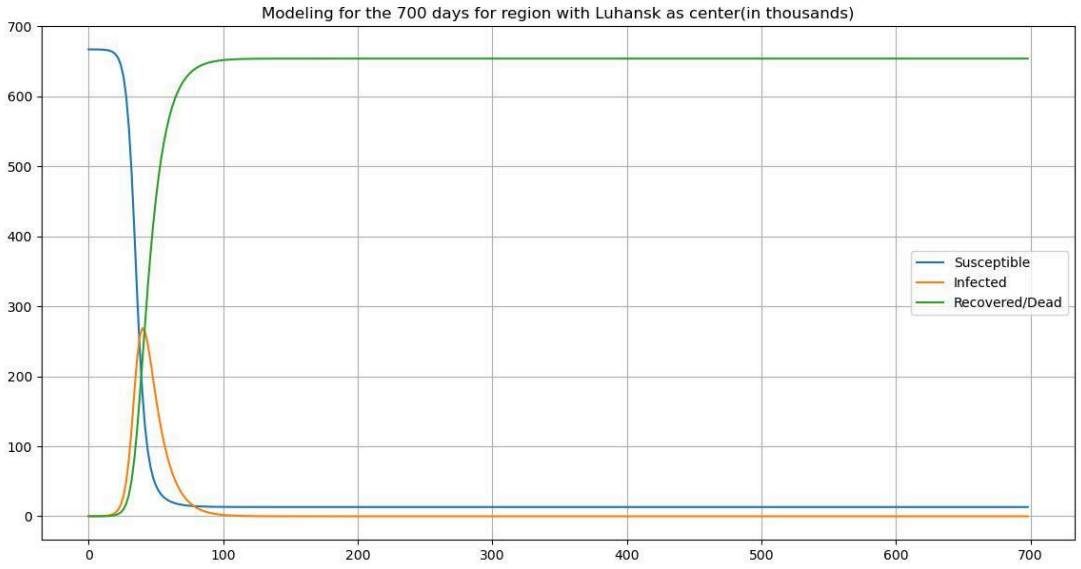

Рис 3.16 Результати моделювання для Луганської області

Для розгляду прикладів результатів моделювання, зображених на рис. 3.14 - 3.16, були вибрані 3 області України, які різняться кількістю населення, а саме Дніпропетровська область(населення майже 3 млн людей), Київська область(населення майже 1.8 млн людей) та Луганська область(населення майже 0,7 млн людей). Як можемо бачити, незалежно від початкової кількості населення(або людей придатних до захворювання), найбільший приріст осіб що захворіли відбувається в перші 100 днів моделювання, після цього, кількість людей що може заразитися та кількість інфікованих спадають до 0, переходячи в стан одужавші/померлі, кількість яких прямує до початкової кількості людей, що можуть заразитися. Для решти областей, результати моделювання для яких зображені на рис. А.1.1 - А.1.22 Додатку А, описана вище динаміка зберігається.

Для кращого розуміння стану захворюваності в кожній з областей, в порівнянні з іншими областями, давайте проілюструємо максимальний відсоток заражених людей в кожній з областей за весь період моделювання відносно загальної кількості населення в області на рис. 3.17:

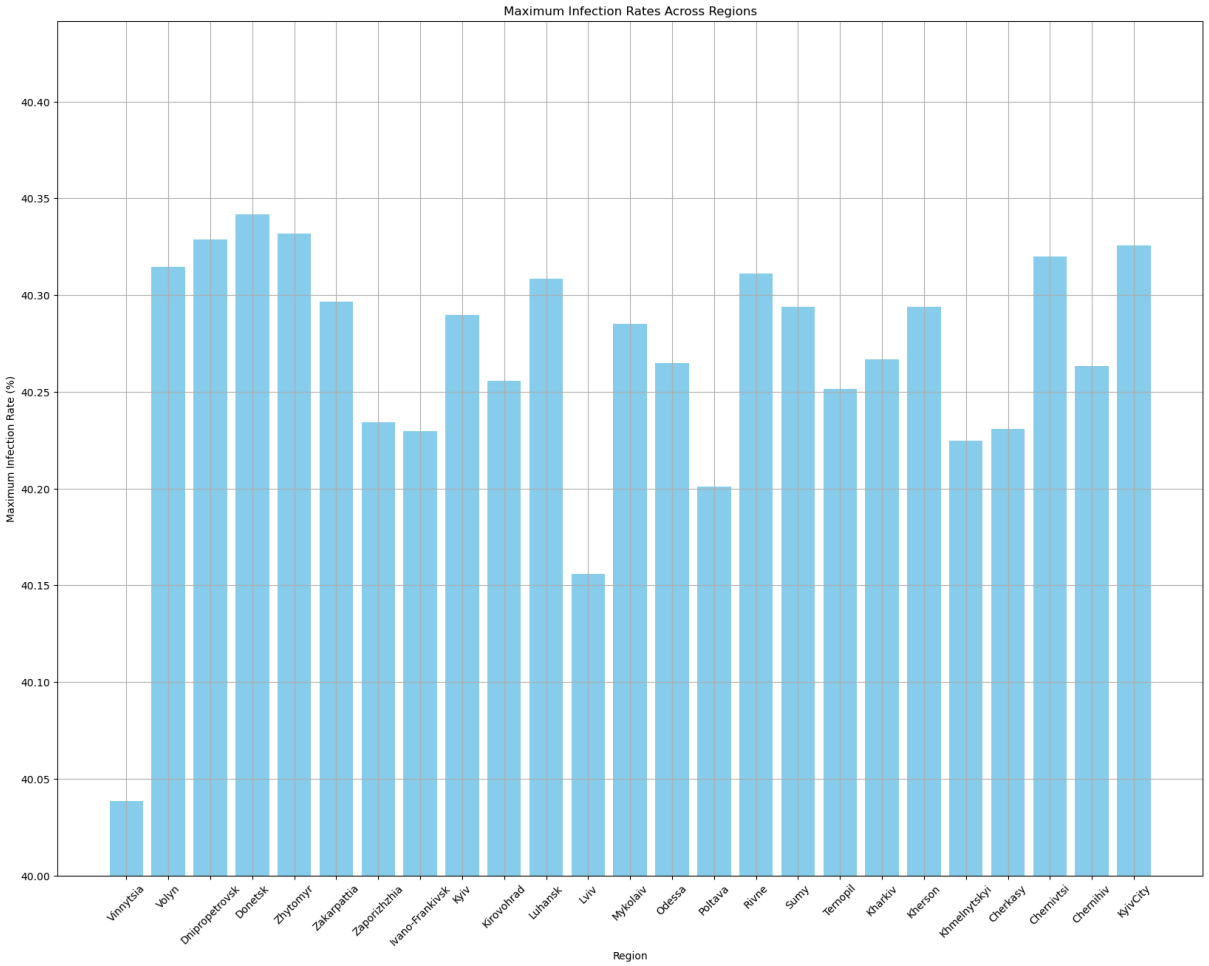

Рис. 3.17 Порівняльна гістограма максимальної кількості заражених відносно загальної кількості населення області

З рис. 3.17 видно, що максимальна кількість заражених індивідуумів відносно загальної кількості населення області за весь проміжок часу моделювання була приблизно однакова, і коливалася навколо 40%. Областю з найменшим значенням максимальної кількості заражених є Вінницька, а з найбільшим - Донецька.

Висновок до розділу

В процесі опису проведених експериментів, ми показали повну роботу алгоритмів Random Forest, починаючи із завантаження наборів даних, їх попередньої обробки та самої роботи алгоритму, разом з проведенням чисельних та візуальних оцінок результатів.

Підсумовуючи висновки роботи моделі Random Forest, можна зазначити,

спираючись на результати значень метрик, MSE та R^2, що дорівнюють орієнтовно 0.0028 та 0.9204 відповідно, що модель досить добре відпрацьовує, та надає достатньо точні передбачення. В умовах реального світу, за наявності більшої деталізації розбиття вікових груп та збільшення загального розміру набору даних, модель може показувати не такі гарні результати передбачень. Проте, для загального розуміння динаміки поширення захворюваності серед населення, для проведення запобіжних заходів, модель відпрацьовує достатньо добре.

Підбиваючи підсумки результатів моделювання за допомогою алгоритму SIR, видно загальну тенденцію поширення захворюваності серед населення областей, і якщо вірити прогнозам моделювання, то вона плачевна. Хоча, варто зазначити, що для моделювання була вибрана найпростіша модель клітинних автоматів, яка не враховує факторів того, чи може людина знову стати заразною після одужання, чи вплив інших захворювань, таких як ВІЛ чи СНІД на розвиток туберкульозу. Також, параметри моделі β і μ потребують кращого тюнінгу, для чого необхідно більше інформації про різні фактори, які впливають на ці параметри. Як показує порівняльна гістограма на рис. 3.17, мінімум 39% населення в певний час будуть заражені туберкульозом, що в дійсності маловірогідно, завдяки заходам, які впроваджують для профілактики та лікування туберкульозу. Отож, найпростіша модель клітинних автоматів SIR не виправдала себе для задачі більш точного моделювання поширення захворювань, але достатньо ефективна, щоб отримати інформацію про загальну тенденцію, тож в подальшому необхідно вибирати більш складний алгоритм, який враховує описані вище недоліки.

Розділ 4. Обговорення результатів дослідження

В розділі 3 були описані робота алгоритмів, а також продемонстровані їх результати. Важливо не тільки оцінити результати наочно, але й за допомогою метрик. Як бачимо з пункту 3.2, для алгоритму Random Forest була проведена оцінка метрик MSE та R^2. За ідеальних умов значення метрики MSE має наближатися до 0, в той час як значення метрики R^2 має наближатися до одиниці.

Значення метрик для результатів роботи алгоритму Random Forest

Таблиця 4.1

Метрика	Значення метрики
MSE	0.0028
R^2	0.9204

Як можемо бачити з табл. 4.1, реальні значення метрик досить наближені до значень метрик в ідеальних умовах, і передбачення, надане даним алгоритмом, є достатньо точним, про що можна сказати при статистичному аналізі набору даних.

Робота алгоритму SIR також була описана в пункті 3.3, після чого, порахувавши результати окремо по кожній з областей та зробивши візуалізацію максимальної кількості заражених індивідуумів відносно загальної кількості населення області за весь проміжок часу моделювання на рис 3.17, ми змогли пересвідчитись в недостатній точності моделювання поширення захворювання, адже середнє значення для всіх областей на рис. 3.17 досягає 39%, що в дійсності не може бути правдою. Такі неточності результатів можна пояснити тим, що не враховуються фактори ризику захворювання, можливості переходити від стану одужання знову до стану зараження та деякі інші. Важливими також в даному алгоритмі виступають параметри, які враховуються при переході між станами здорові-інфіковані та інфіковані-одужавші/померлі. Ці параметри потребують кращого підбору, для чого необхідна більша наукова база.

Загалом, порівнюючи вибрані алгоритми з тими, що були вибрані іншими

дослідниками в суміжних темах, можна чітко побачити, що вибрані моделі є архітектурно найпростішими в своїй області, і звичайно, їх простота і забирає в них високу точність надання результатів.

Для подальшого аналізу даної теми, варто вибрати більш структурно складні алгоритми, точність передбачень яких буде більше, а саме рекурсивні нейронні мережі, або прогнозування за допомогою часових рядів.

Висновки до розділу

З розділу "Обговорення результатів" видно, що обрані алгоритми Random Forest і SIR були детально досліджені та проаналізовані з використанням відповідних метрик. Для Random Forest проведена оцінка за метриками MSE і R2, які наближаються до ідеальних значень, свідчать про точність передбачень цього алгоритму.

У випадку моделювання SIR було виявлено недостатню точність у передбаченні поширення захворювання. Зокрема, візуалізація максимальної кількості заражених відносно загальної кількості населення області показала середнє значення навколо 39%, що виявилося не реалістичним. Ці неточності можна пояснити недоліками моделі, такими як неврахування факторів ризику захворювання та можливість повторного зараження після одужання.

Порівнюючи вибрані алгоритми з іншими дослідженнями в суміжних темах, можна зрозуміти, що вони є архітектурно найпростішими в своїй області. Простота цих моделей забирає можливість надання результатів високої точності, проте для подальшого аналізу теми рекомендується розглянути більш архітектурно складні алгоритми, такі як рекурсивні нейронні мережі або прогнозування за допомогою часових рядів, які можуть забезпечити ще більшу точність передбачень.

Розділ. 5 Висновки

В ході курсової роботи було проведено дослідження поширення туберкульозу та прогнозування найбільш уразливих соціальних груп населення. В розділі Вступ чітко сформульовано мету дослідження, а також проведено аналіз актуальності обраної проблеми та визначено об'єкт та предмет дослідження. У Розділі 1 проведено аналіз літературних джерел та вибір методів для моделювання поширення захворювання, в результаті чого були обрані методи Random Forest та SIR.

У Розділі 2 були проведені попередні роботи з аналізу даних та обрано найбільш підходящий набір даних для дослідження. Також, проведено порівняльний аналіз методів, використаних іншими дослідниками, і вибрано методи для подальшого дослідження.

У Розділі 3 були детально описані та проаналізовані результати застосування методів Random Forest і SIR. Для Random Forest були оцінені метрики MSE та R2, що показали високу точність передбачень. У випадку моделювання алгоритмом SIR, за допомогою візуальної оцінки результатів, було виявлено недостатню точність, що обумовлено недоліками моделі. Порівнюючи обрані методи з іншими дослідженнями, було зроблено висновок про необхідність розгляду більш складних алгоритмів для отримання більш точних результатів.

Отже, на основі результатів дослідження можна зробити висновок про достатню ефективність методу Random Forest для та прогнозування уразливих соціальних груп населення та слабку ефективність алгоритму SIR для моделювання поширення туберкульозу. Для подальшого розвитку дослідження рекомендується розгляд більш складних алгоритмів та врахування додаткових факторів, що впливають на поширення захворювання. Крім того, для кращого розуміння подальших дій для поротьби з хворобою, доцільно буде провести симуляцію поширення туберкульозу серед населення України.

Джерела літератури

[1] Ільницький Г.І. Математичне моделювання епідеміологічних та клініколабораторних проявів туберкульозу. Доступний:https://lpnu.ua/sites/default/files/2021/radaphd/9119/ilnickiy-gi-dis.pdf

[2] Туберкульоз В Україні. Аналітично-статистичний довідник за 2021 р. Доступний:https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiFy6WjrtiEAxVy9wIHHTkrBeEQFnoECA8QAQ&url=https%3A%2F%2Fphc.org.ua%2Fsites%2Fdefault%2Ffiles%2Fusers%2Fuser90%2FTB_surveillance_statistical-information_2021_dovidnyk.docx&usg=AOvVaw3IhYDYNB5iE3viqs0PPTV1&opi=89978449

[3] Роговський М. Г. Математичні моделі, що описують процеси розповсюдження захворювань. Доступний:https://ela.kpi.ua/bitstream/123456789/29479/1/Rohovskyi_bakalavr.docx

[4] Шевченко Я. М. Використання біологічного моделювання для з'ясування ролі генів у функціонуванні організмів. Доступний:http://ekhsuir.kspu.edu/bitstream/handle/123456789/16815/Шевченко%20Ярослава_211М_Дипломна%20Робота.pdf?sequence=1&isAllowed=y

[5] L.D. Todoriko & others. Prospects for the use of artificial intelligence to predict the spread of tuberculosis infection in the WHO European Region. Доступний:http://tubvil.com.ua/article/view/282415/276650

[6] Хімія, екологія і освіта Доступний:https://reposit.nupp.edu.ua/bitstream/PoltNTU/7617/3/IV%20%20МНПК_71.pdf

[7] Ukraine. Доступний:https://drive.google.com/drive/folders/1RpivIgIFGlvTCNRJXSwY3A0-FRySJHvg

[8] Sklearn Documentation. [Онлайновий]. Доступний:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[9] Seaborn Documentation. [Онлайновий]. Доступний:https://seaborn.pydata.org/generated/seaborn.scatterplot.html

[10] Matplotlib Documentation. [Онлайновий]. Доступний:https://matplotlib.org

[11] M. Kröger, R. Schlickeiser. Verification of the accuracy of the SIR model in forecasting based on the improved SIR model with a constant ratio of recovery to infection rate by comparing with monitored second wave data. [Онлайновий]. Доступний:https://royalsocietypublishing.org/doi/10.1098/rsos.211379

[12] The SIR model studies the population of susceptible(S), infectious(I), and recovered or removed (R) over time, often utilizing ODE. [Онлайновий]. Доступний:https://www.sciencedirect.com/topics/mathematics/sir-model

ДОДАТОК А

А.1 Результати моделювання за допомогою алгоритму SIR для областей України, які не були згадані в пункті 3.3, та міста Київ.

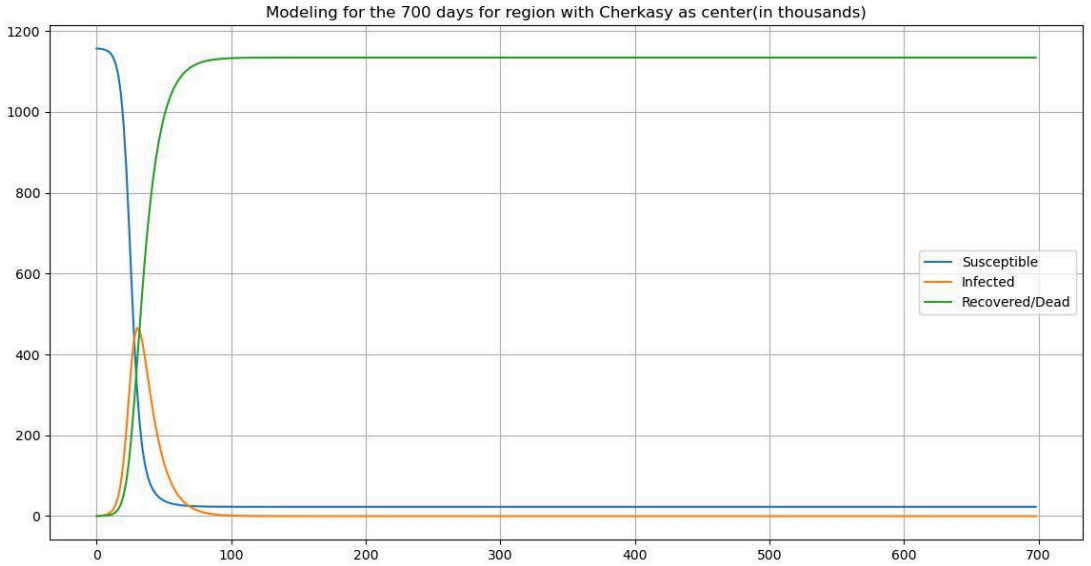

Рис А.1.1 Результати моделювання для Черкаської області

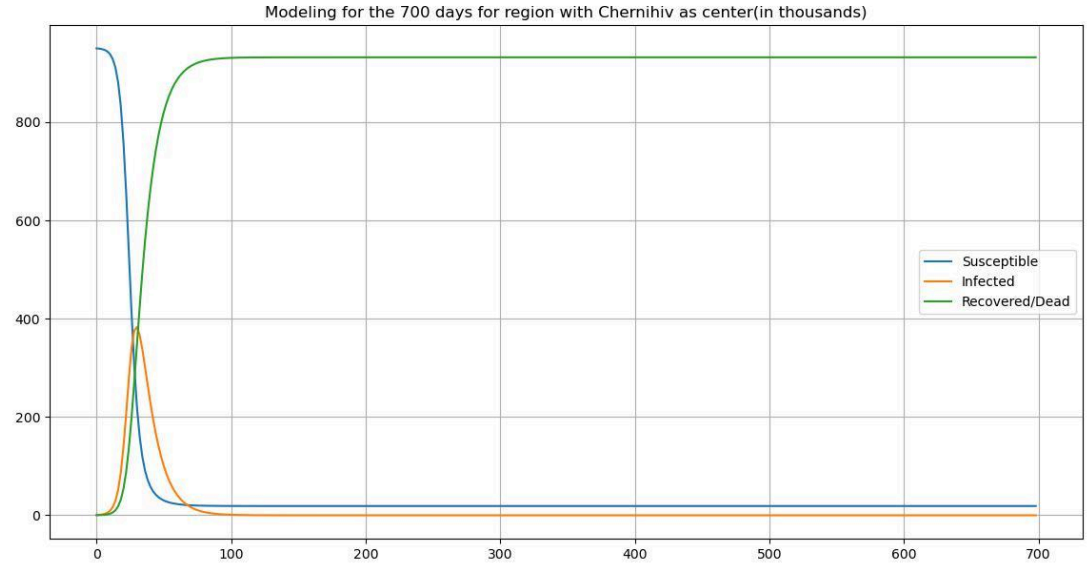

Рис А.1.2 Результати моделювання для Чернігівської області

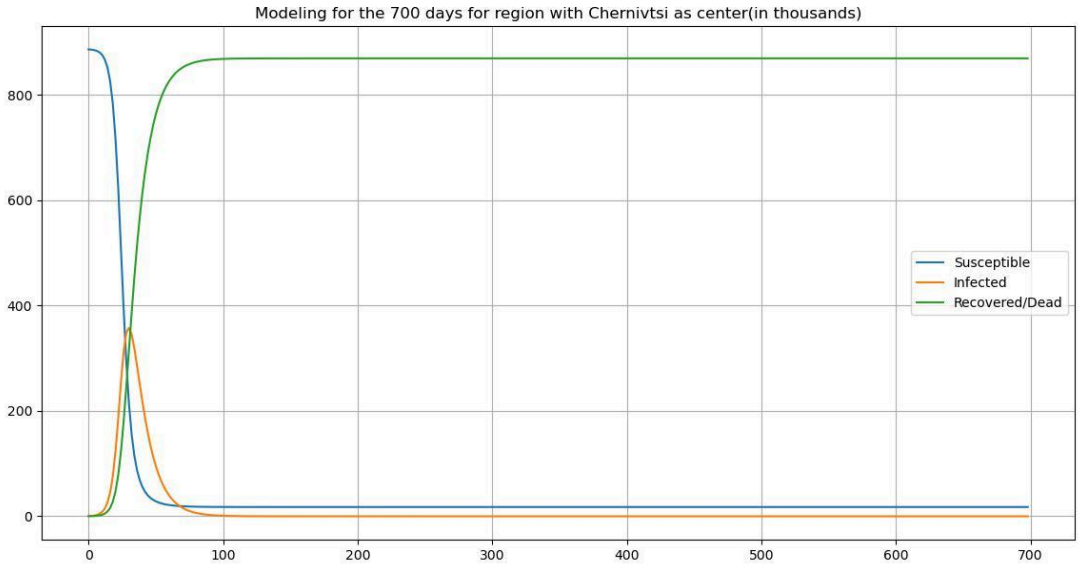

Рис А.1.3 Результати моделювання для Чернівецької області

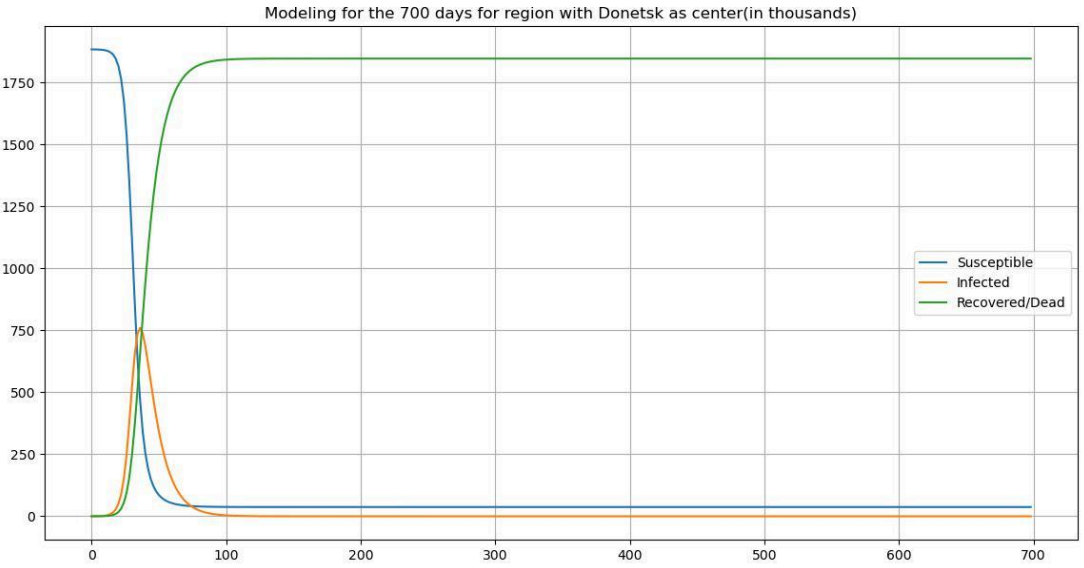

Рис А.1.4 Результати моделювання для Донецької області

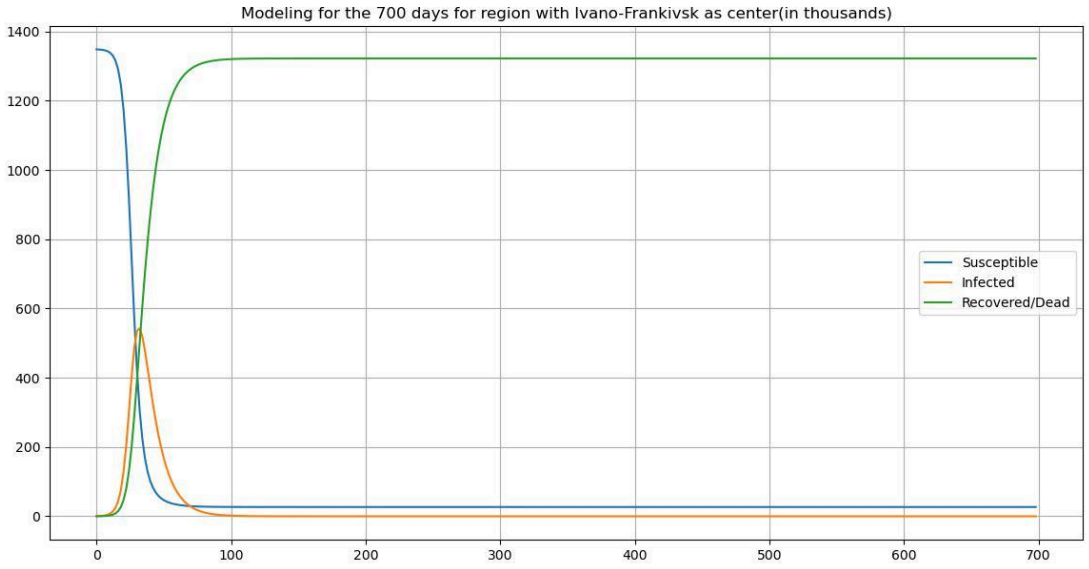

Рис А.1.5 Результати моделювання для Івано-Франківської області

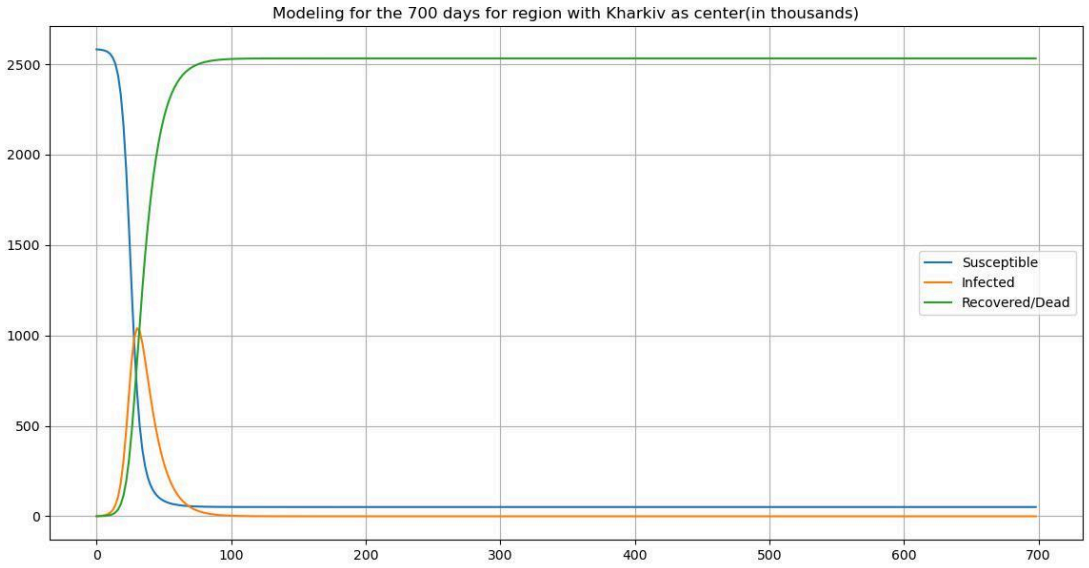

Рис А.1.6 Результати моделювання для Харківської області

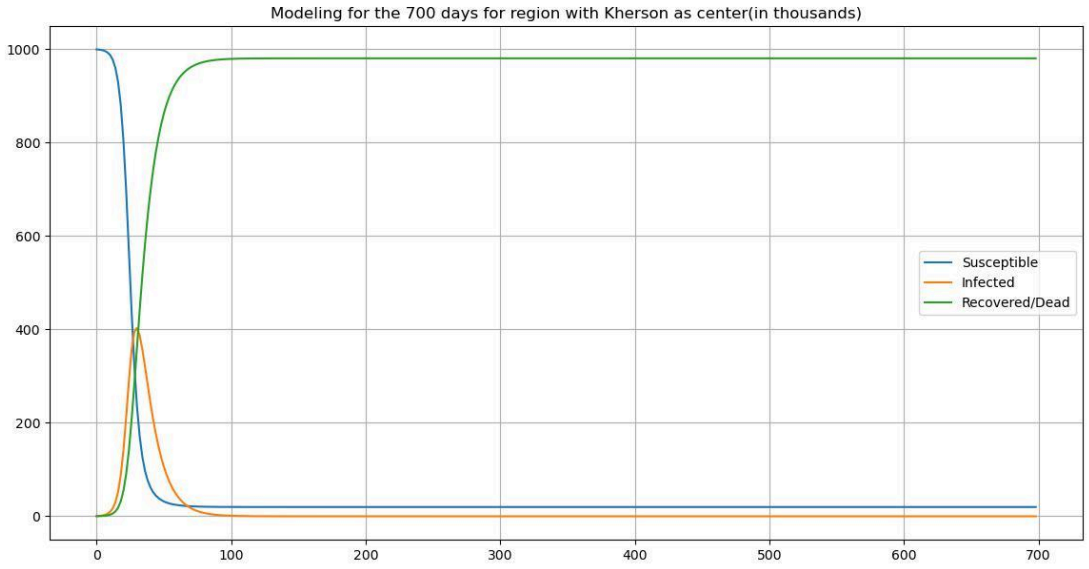

Рис А.1.7 Результати моделювання для Херсонської області

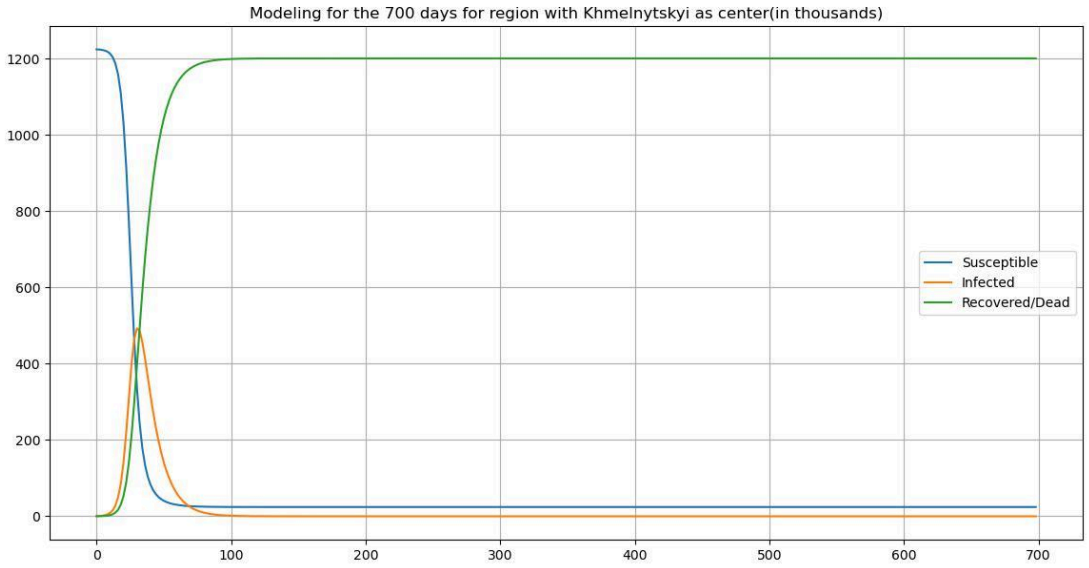

Рис А.1.8 Результати моделювання для Хмельницької області

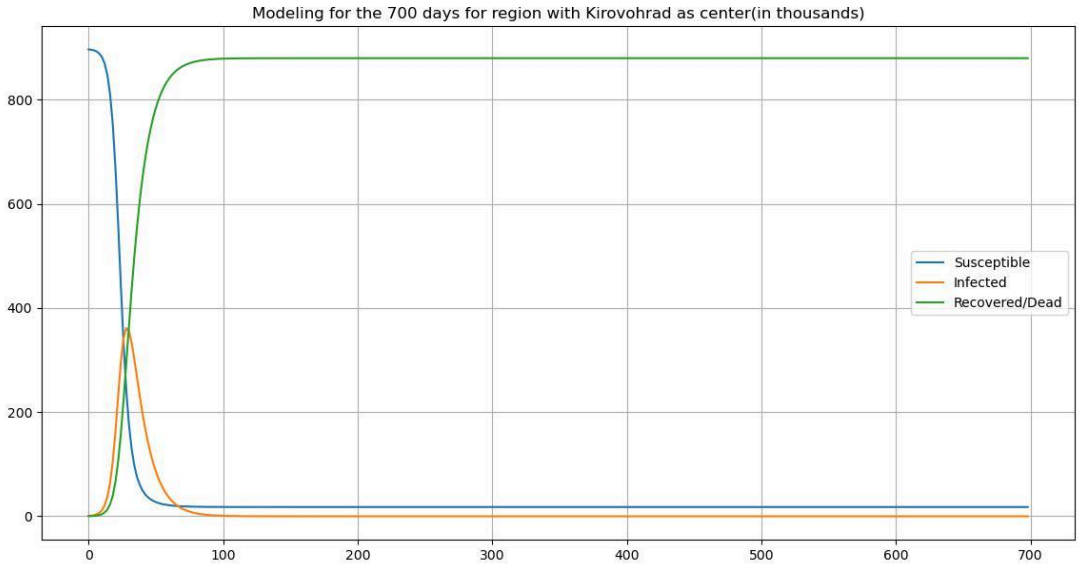

Рис А.1.9 Результати моделювання для Кіровоградської області

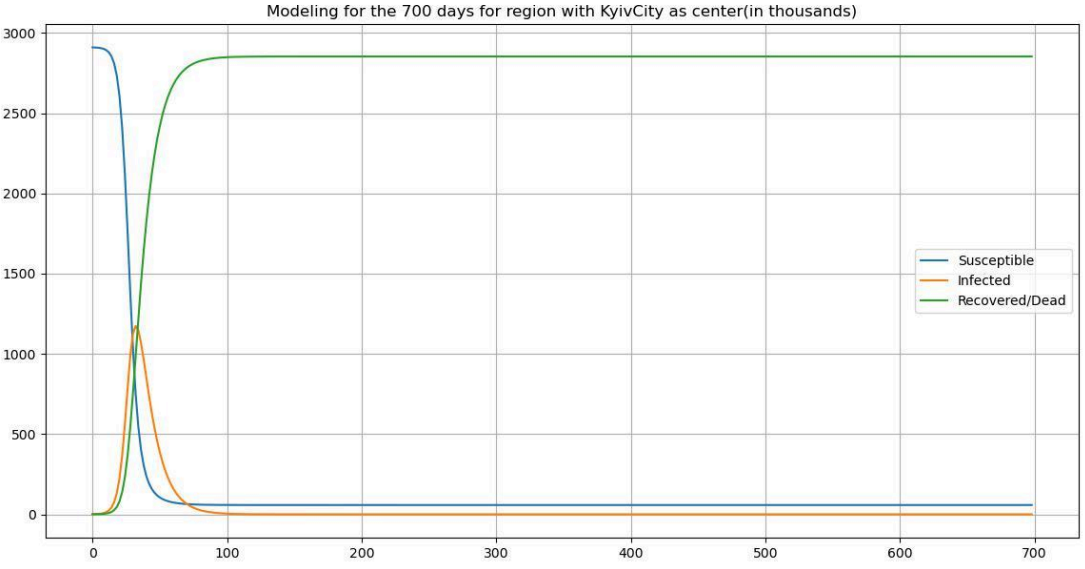

Рис А.1.10 Результати моделювання для міста Київ

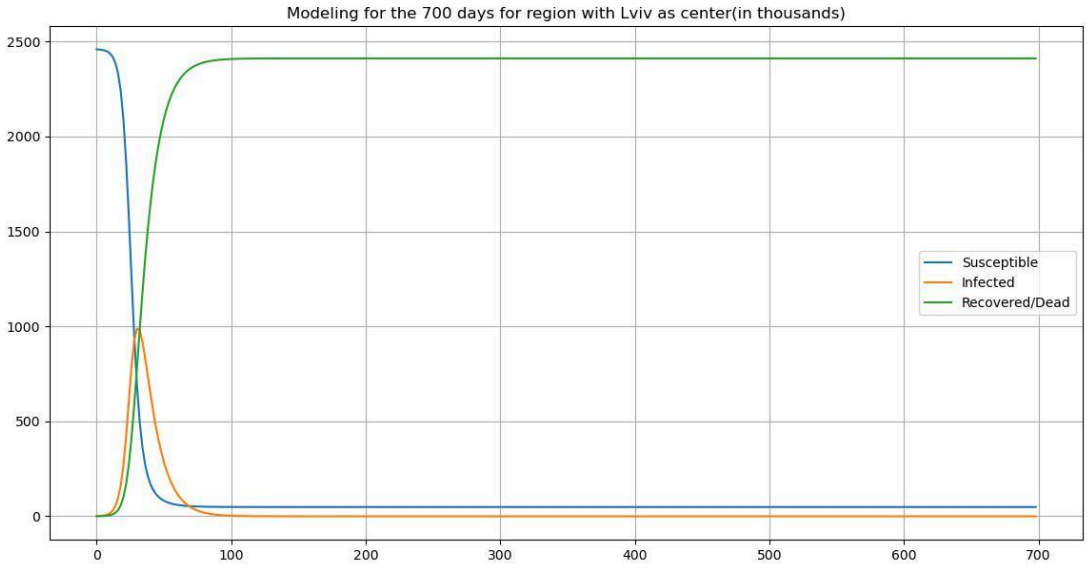

Рис А.1.11 Результати моделювання для Львівської області

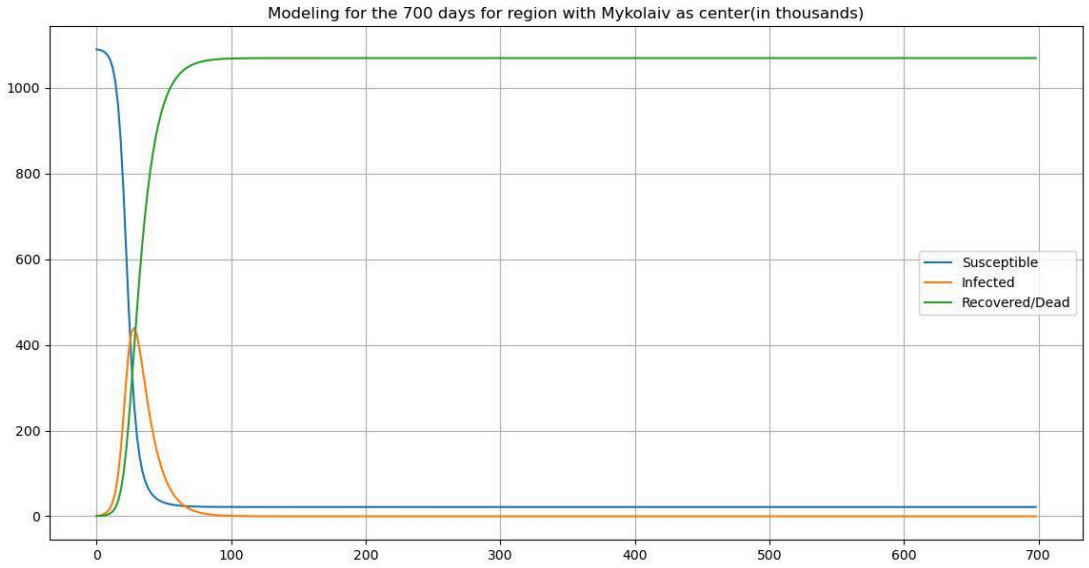

Рис А.1.12 Результати моделювання для Миколаївської області

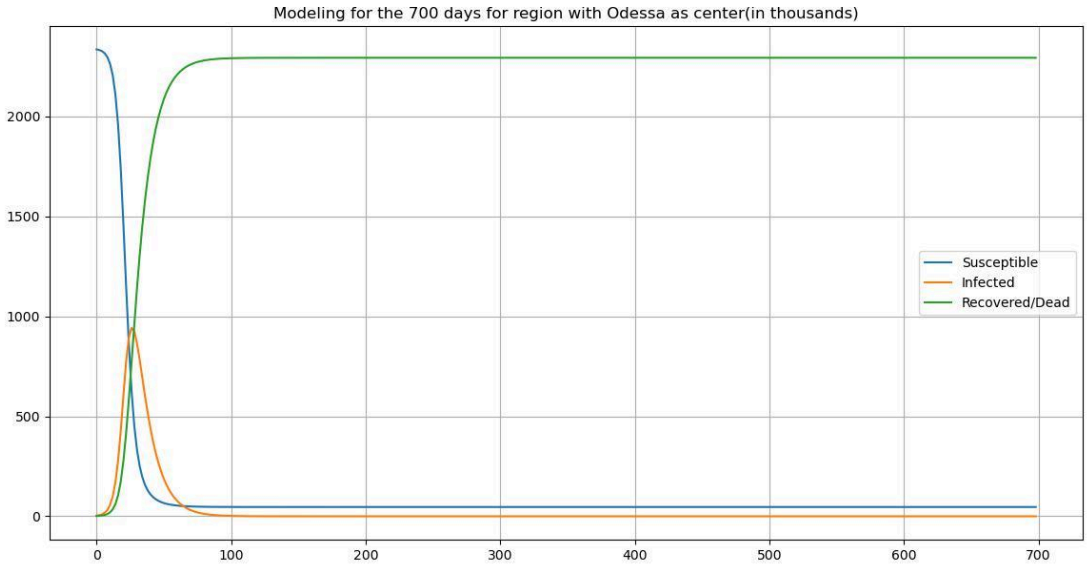

Рис А.1.13 Результати моделювання для Одеської області

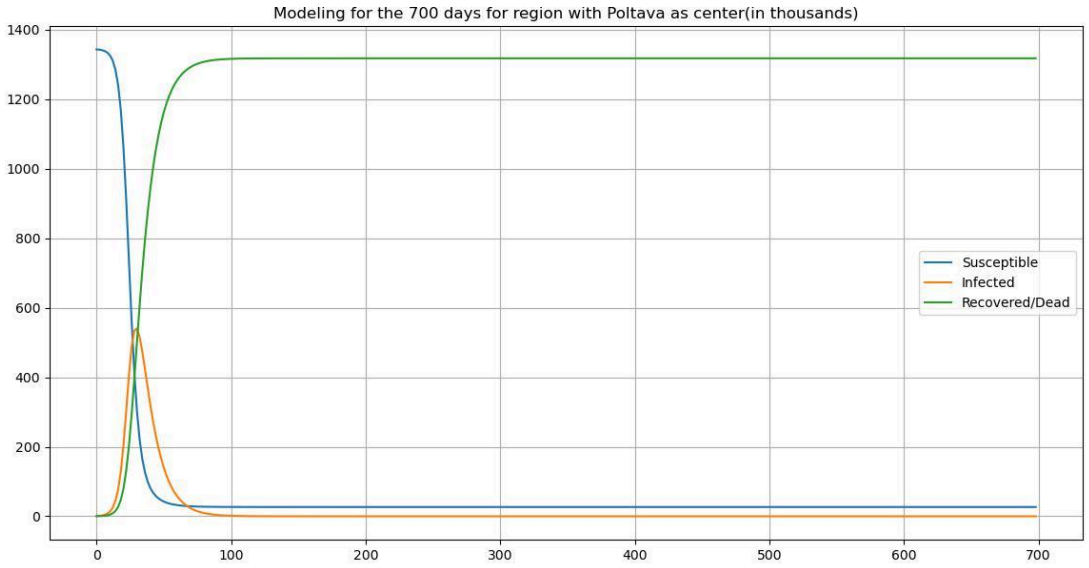

Рис А.1.14 Результати моделювання для Полтавської області

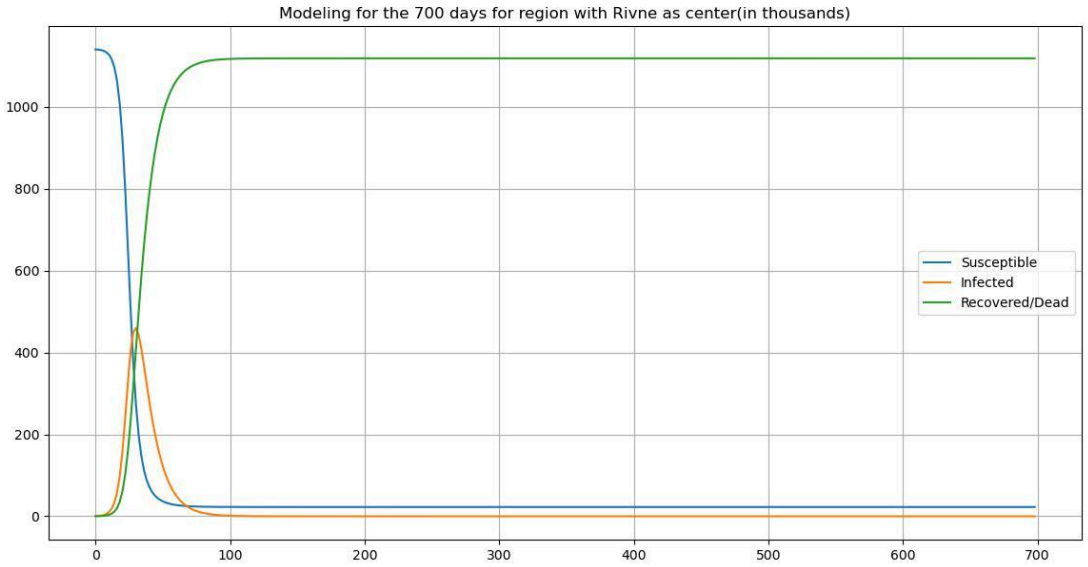

Рис А.1.15 Результати моделювання для Рівненської області

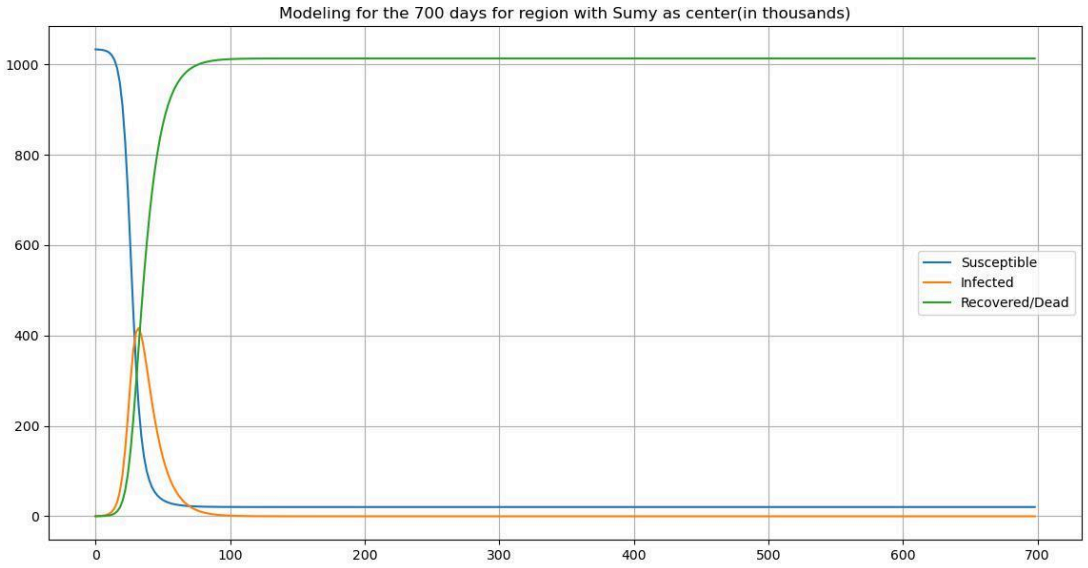

Рис А.1.16 Результати моделювання для Сумської області

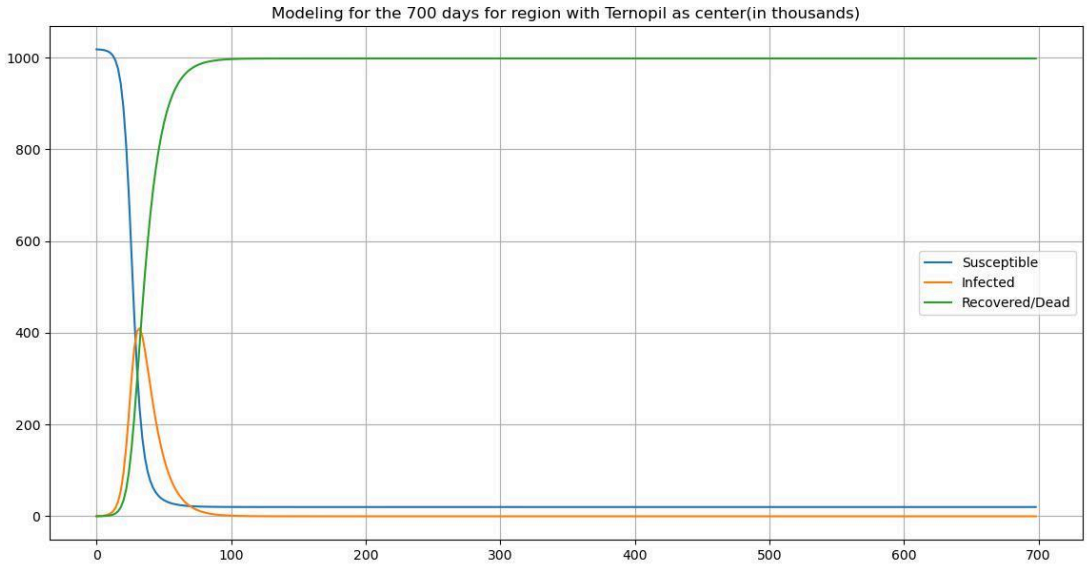

Рис А.1.17 Результати моделювання для Тернопільської області

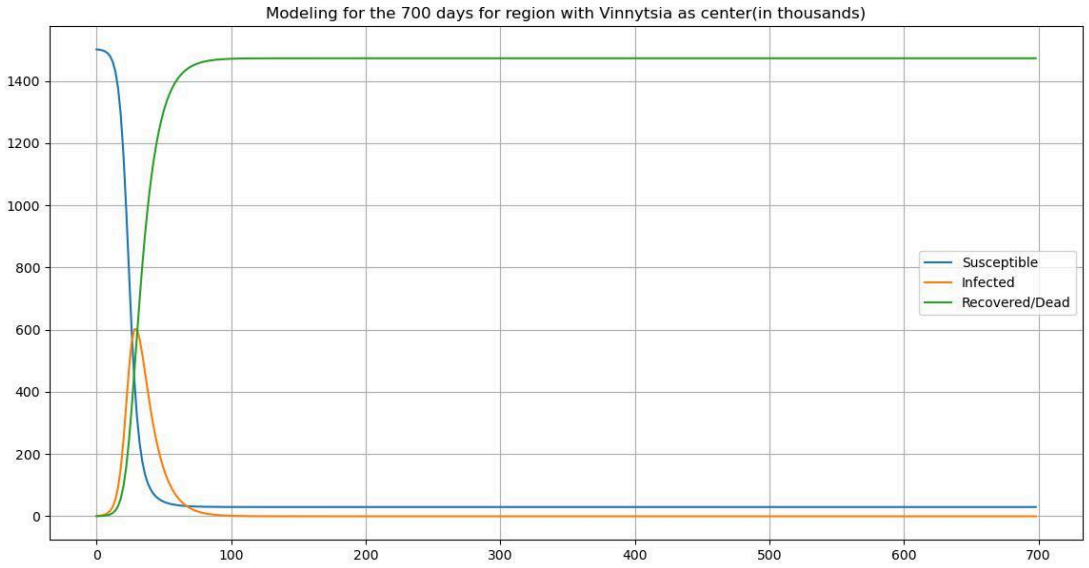

Рис А.1.18 Результати моделювання для Вінницької області

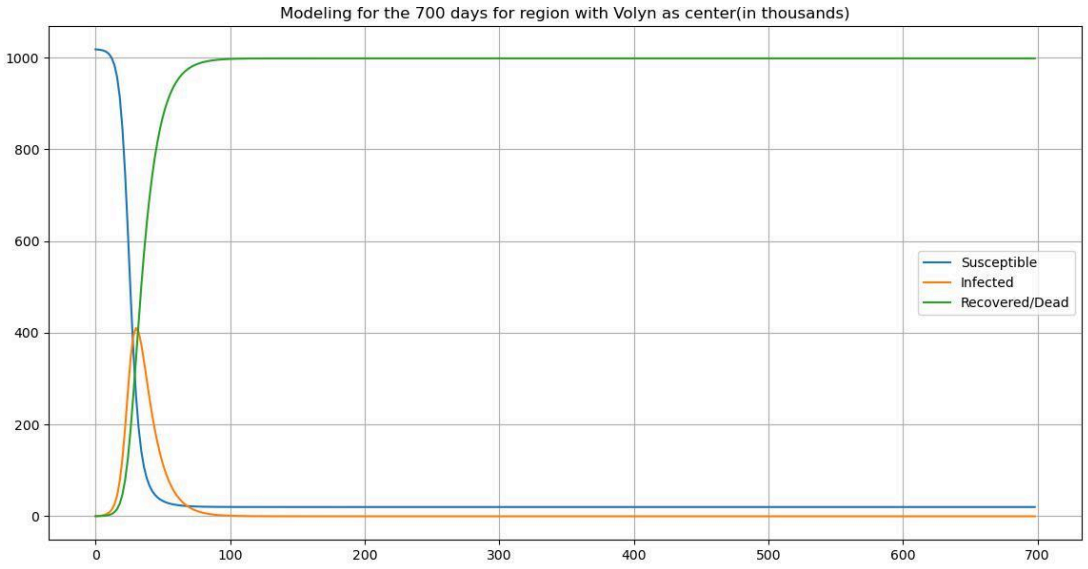

Рис А.1.19 Результати моделювання для Волинської області

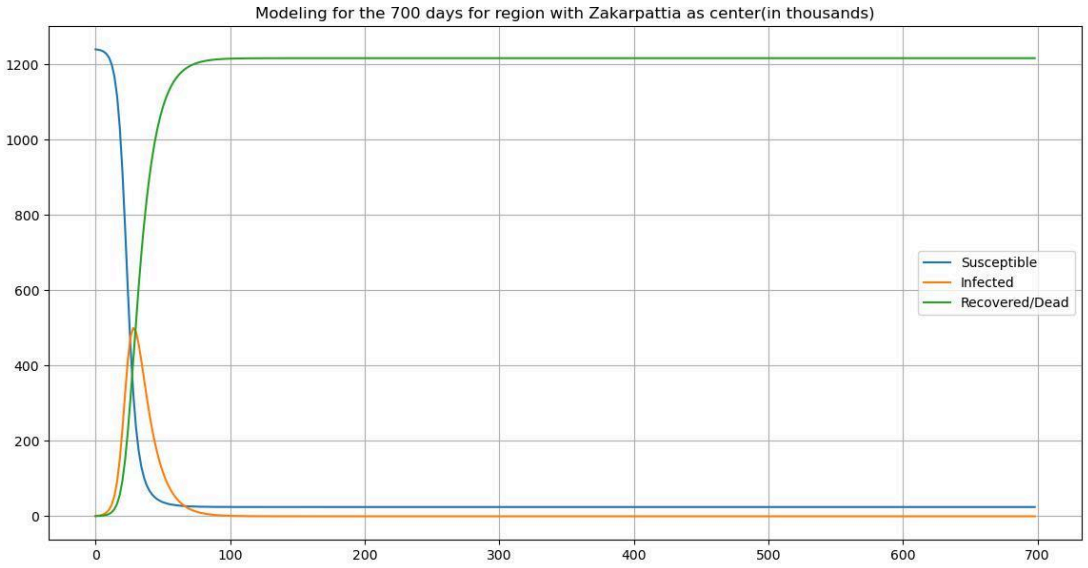

Рис А.1.20 Результати моделювання для Закарпатської області

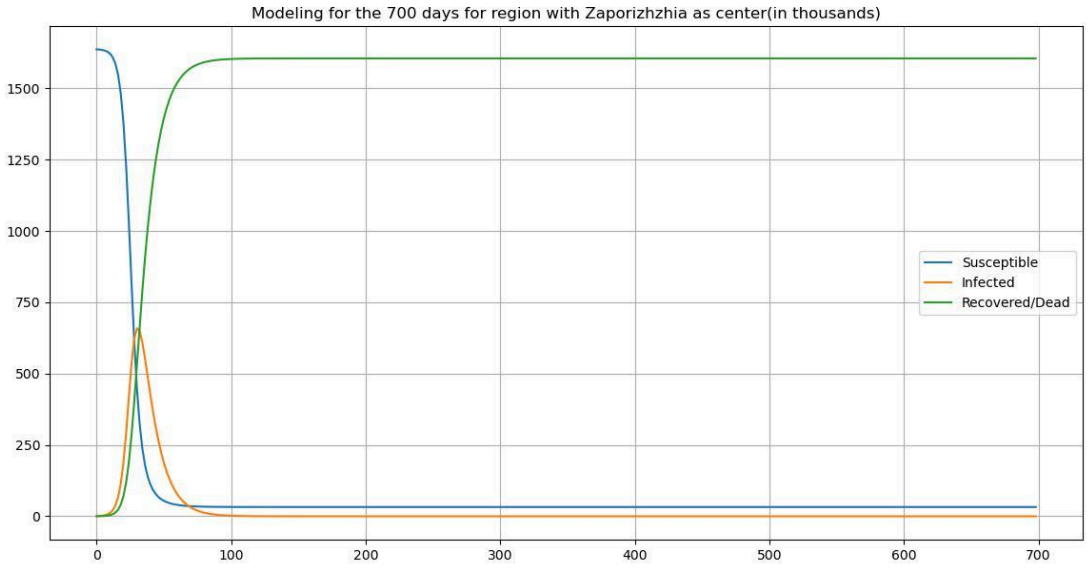

Рис А.1.21 Результати моделювання для Запорізької області

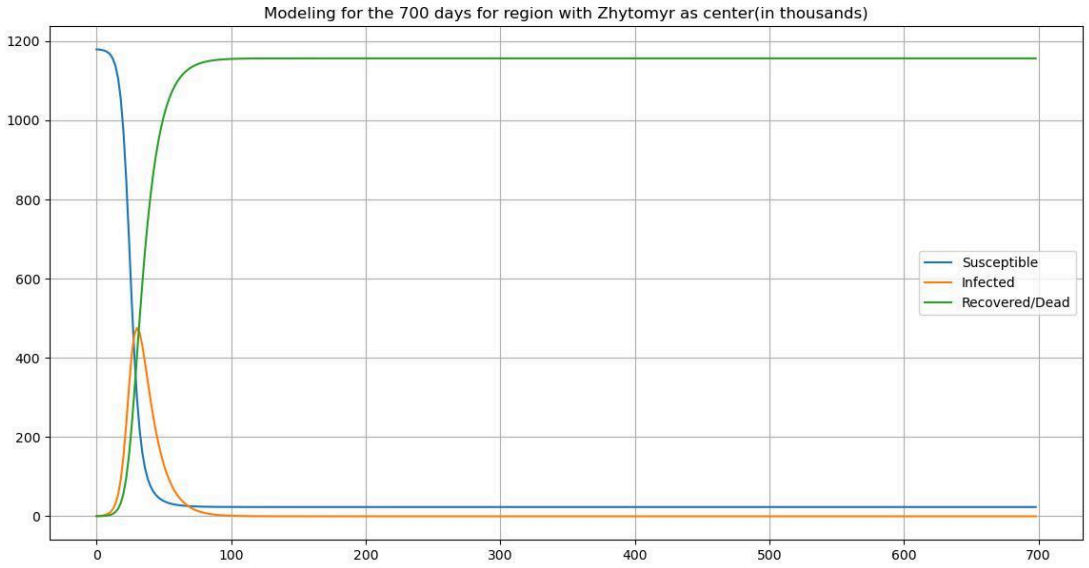

Рис А.1.22 Результати моделювання для Житомирської області області

А. 2 Посилання на репозиторій з кодом до проекту

https://github.com/rDrayBen/CourseWorkML/blob/main/Code/Full_Code.ipynb