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 – 2023

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.

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

Актуальність теми відображає важливість і суттєвість проблеми, яку досліджує автор, 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. Спричинене цим високим рівнем захворюваності та потенційною загрозою

здоров'ю громадського здоров'я це важлива проблема, 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 науковий внесок, 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, може сприяти вдосконаленню методів діагностики та лікування туберкульозу, а також розробці ефективних програм контролю та профілактики.

Об’єкт та предмет дослідження

Об'єктом дослідження є поширення туберкульозу на території областей України. Предметом дослідження є аналіз соціально-демографічних аспектів поширення туберкульозу та визначення найуразливіших груп населення за ознаками статі та віку. 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.

Методи дослідження

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 класифікація 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.

Аналіз літературних джерел

1) Ільницький Г. І. (2021) 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. Модель ґрунтується на системі диференціальних рівнянь, які описують динаміку поширення туберкульозу в популяції. Модель враховує такі фактори, як: рівень захворюваності на туберкульоз, смертність від туберкульозу, рівень інфікування та ефективність лікування. Модель не є дуже підходящою 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.
2) Роговський В. О. (2023) також розробив 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.

3) За даними Центру громадського здоров'я МОЗ України, 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%.
4) У розділі "Створення Байєсівської мережі факторів ризику захворіти COVID-19" дипломної роботи Шевченко Ярослава описує 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).
5) З журналу “Хімія, екологія та освіта”, розділу “МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ПОШИРЕННЯ ВІРУСНИХ ІНФЕКЦІЙ В ЛОКАЛЬНИХ УРБОЕКОСИСТЕМАХ” можна дізнатися, що було досліджено природу 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).

6) 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” тільки це підтверджує. У цій статті 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.

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	Очікування та	Простий метод	Не враховує вплив

громадського здоров'я МОЗ України (2022)	реальні результати щодо зниження захворюваності та смертності від туберкульозу	прогнозування	різних факторів
Шевченко Ярослав (дипломна робота)	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 штучного інтелекту для прогнозування поширення туберкульозної інфекції в Європейському регіоні ВООЗ”	Класична SIR-модель, модель міського середовища	Кращий прогноз, врахування геопросторових даних, швидкість розрахунків, можливість паралельних обчислень	Складність розробки, потреба в багатьох даних

Висновок з аналізу літературних джерел підтверджує актуальність проблеми 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 туберкульозу 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.

Аналіз матеріалів та методів

2.1. Датасет

Для передбачення поширення туберкульозу та його моделювання, необхідним елементом виступає датасет, який буде нести в собі достатню кількість інформації, аби передбачення мали високу точність, саме тому, вибраний датасет[https://drive.google.com/drive/folders/1RpivIgIFGlvTCNRJXSwY3A0-FRySJHvg] містить дані про захворюваність на туберкульоз в Україні за тривалий період 2007- 2022 роки, поділені на часові проміжки. 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, 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.

2.2. Дерево прийняття рішень(ДПР)(класифікатор)

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 даними середнього розміру, що саме підпадає під наш випадок.

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 області, кількості захворювання по конкретній області, вікових категорій та статей захворюваних. У результаті опрацювання даних, модель буде подавати на вихід вікову категорію і стать людей по кожній з областей, 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.

Алгоритм роботи Random Forest:

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

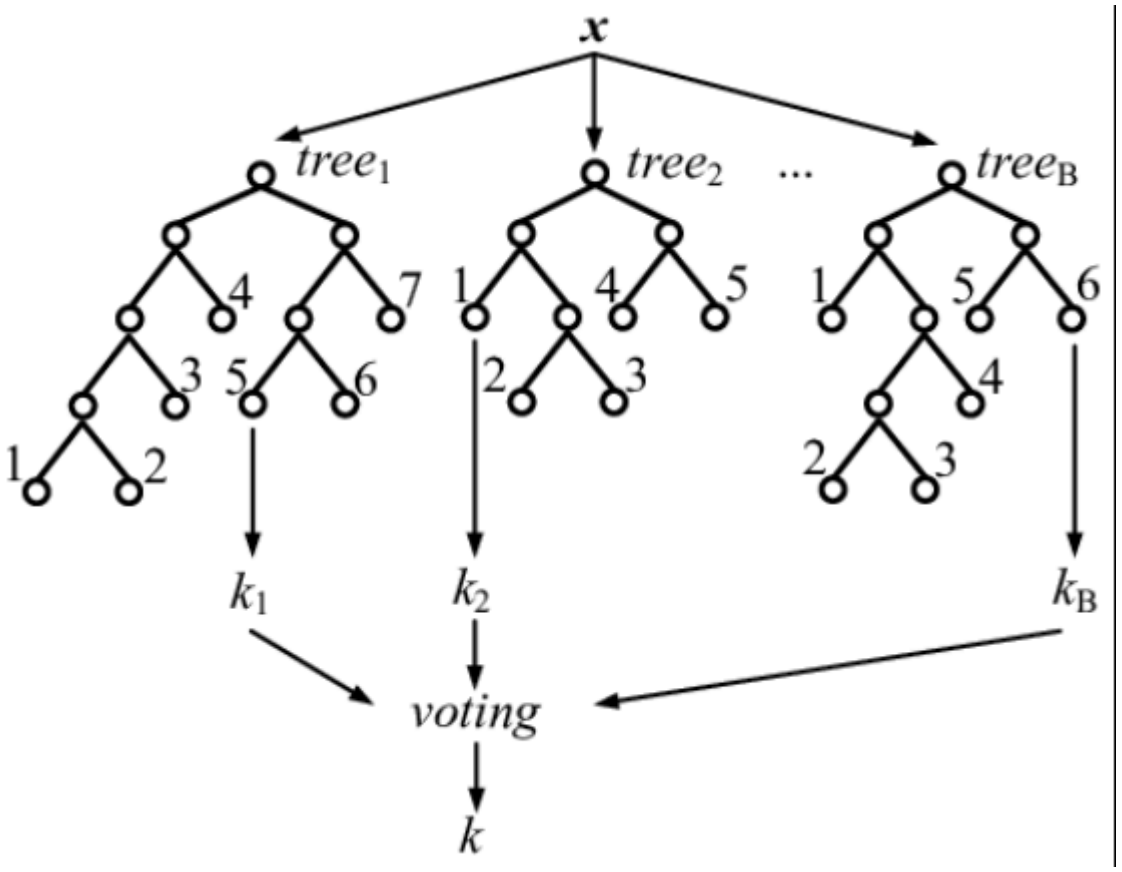
    для кожного дерева:

    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

Для оцінки точності можна використати F1-міру(зважене середнє значення точності передбачення), а для оцінки похибки моделі середньоквадратичну похибка(середнє значення квадратів різниць між прогнозами та дійсними значеннями).



Мал. 2.2. Умовна архітектура Random Forest

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 КА, 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.

Звичайно, результати отримані завдяки моделі SIR, будуть менш точними за результати отримані завдяки Random Forest, але завдяки порівнянню результатів, ми також зможемо визначити яка з методик краще підходить для вирішення даного питання.

Алгоритм роботи моделі SIR:

1) Ініціалізація:
    a) Встановимо сітку клітин.
    b) Виберемо початкові значення для S, I та R.
2) Оновлення(для кожної клітини):
    a) З ймовірністю β інфікована клітина заражає сприятливу клітину.
    b) З ймовірністю γ інфікована клітина одужує.
3) Повторення:
    a) Повторюємо кроки 1 і 2, доки не буде досягнуто кінцевого стану.

Для оцінки похибки моделі використаємо середньоквадратичну похибку(середнє значення квадратів різниць між прогнозами та дійсними значеннями).

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).