МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**Національний університет «Запорізька політехніка»**

Факультет «Комп'ютерних наук і технологій»

Кафедра «Системний аналіз та обчислювальна математика»

# Пояснювальна записка

до дипломного проєкту (роботи)

бакалавра

на тему: «ML моделі впливу якості повітря на рівень захворюваності»

Виконав Студент 4 курсу, групи КНТ-811

Спеціальності 124 «Системний аналіз»

Освітня програма (спеціалізація)

інтелектуальні технології та

прийняття рішень в складних системах

ПОГОДАЄВ Д.В.

Керівник ШИРОКОРАД Д.В.

Рецензент ДУМІН О.М.

2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**Національний університет «Запорізька політехніка»**

Факультет «Комп'ютерних наук і технологій»

Кафедра «Системний аналіз та обчислювальна математика»

Ступінь вищою освіти бакалавр

Спеціальність 124 «Системний аналіз»

Освітня програма (спеціалізація) інтелектуальні технології та прийняття рішень в складних системах

**ЗАТВЕРДЖУЮ**

**Завідувач кафедри**: к.ф.-м.н.,

доцент ТЕРЕЩЕНКО Е.В.

«___»_____ 2025 року

# ЗАВДАННЯ

НА ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ) СТУДЕНТА

ПОГОДАЄВА Дмитра Володимировича

1. Тема проєкту (роботи) ML моделі впливу якості повітря на рівень захворюваності.

керівник проєкту (роботи) к.ф.-м.н. доцент ШИРОКОРАД Д.В.

затверджені наказом закладу вищої освіти від «16» травня 2025 року № 265

2. Строк подання студентом проєкту (роботи) «16» червня 2025 року

3. Вихідні дані до проєкту (роботи) Офіційна статистика, відкриті бази даних, демографічні показники та медичні дані щодо захворюваності населення.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) У першому розділі проведено аналіз існуючих досліджень економіки, впливу війни на довкілля та здоров'я населення. У другому розділі описані джерела даних: показники якості повітря, тягар хвороб та економічні параметри. Третій розділ присвячений підготовці даних до аналізу, а у четвертому розглянуто побудову нейромережі для моделювання впливу стану економіки та якості повітря на захворюваність.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, кількість слайдів, плакатів)

6. Консультанти розділів проєкту (роботи)

| Розділ | ПРІЗВИЩЕ, ініціали та посада консультанта | Підпис, дата | |
|---|---|---|---|
| | | завдання видав | Прийняв виконане завдання |
| 1 | ШИРОКОРАД Д.В., к.ф.- м.н., доцент | 1.02.2025 | 28.02.2025 |
| 2 | ШИРОКОРАД Д.В., к.ф.- м.н., доцент | 1.03.2025 | 30.03.2025 |
| 3 | ШИРОКОРАД Д.В., к.ф.- м.н., доцент | 1.04.2025 | 30.04.2025 |
| 4 | ШИРОКОРАД Д.В., к.ф.- м.н., доцент | 1.05.2025 | 30.05.2025 |
| Нормоконтроль | ШИРОКОРАД Д.В., к.ф.- м.н., доцент | 1.06.2025 | 6.05.2025 |

7. Дата видачі завдання «_____» _____ 2025 року.

## КАЛЕНДАРНИЙ ПЛАН

| № з/п | Назви етапів роботи | Термін виконання етапів роботи |
|---|---|---|
| 1. | Сформулювати мету та основні завдання дипломної роботи | Січень 2025 |
| 2. | Опрацювати літературу та існуючі дослідження за темою роботи | Січень 2025 |
| 3. | Розробка програмної реалізації для вирішення задачі | Лютий 2025 |
| 4. | Розрахунки та аналіз даних | Березень-квітень 2025 |
| 5. | Оформлення пояснювальної записки | Травень 2025 |
| 6. | Попередній захист дипломної роботи та отримання рецензій | 04.06.2025-15.06.2025 |
| 7. | Захист дипломної роботи | 16.06.2025 |

**Студент(ка)** _____ ПОГОДАЄВ Д.В.

( підпис )

**Керівник проєкту (роботи)** _____ ШИРОКОРАД Д.В.

( підпис )

# РЕФЕРАТ

Наукова робота: 62 сторінки, 4 таблиці, 31 рисунок, 6 додатків, 28 джерел.

Мета роботи полягає в розробці та впровадженні нейромережевої моделі для прогнозування захворюваності населення України. Це дозволить забезпечити точний аналіз динаміки поширення захворювань, виявити тренди та аномалії в епідеміологічних даних, а також оптимізувати заходи для попередження та контролю захворювань на основі отриманих прогнозів.

Наукова новизна полягає у використанні сучасних нейромережевих методів до аналізу відкритих даних України з урахуванням специфіки епідеміологічних даних регіону. Запропонована модель забезпечує високу точність прогнозування завдяки адаптації до динамічних змін у поширенні захворювань, а також врахуванню різних чинників ризику та сезонності. Такий підхід дозволяє вперше отримати прогнозні дані, які можуть бути використані для ефективного планування та оптимізації ресурсів у сфері охорони здоров'я.

Практична цінність отриманих результатів викликана необхідністю розробки ефективних стратегій моніторингу та заходів для зменшення впливу воєнних дій на якість повітря та забезпечення безпеки здоров'я громадян.

Ключові слова: ПОКАЗНИКИ ЯКОСТІ ПОВІТРЯ, ЗАХВОРЮВАНІСТЬ, НЕЙРОННА МЕРЕЖА.

ЗМІСТ

# ВСТУП

Аналіз взаємозв'язку між станом економіки, якістю повітря та захворюваністю населення ускладнюється в умовах сучасних конфліктів, коли військові дії можуть призвести до серйозних наслідків для якості атмосферного повітря та, відповідно, для здоров'я громадян.

Вплив військових дій на стан довкілля: Воєнні дії супроводжуються інтенсивними руйнуваннями, що може призводити до пожеж, вибухів та інших подій, що значно забруднюють повітря. Руйнування інфраструктури може вивільняти шкідливі речовини та забруднюючі частки, що має дійсний вплив на якість атмосферного повітря.

Загроза для здоров'я населення: Забруднене повітря внаслідок військових конфліктів стає фактором, який підвищує ризик респіраторних захворювань, алергій, серцево-судинних захворювань та інших проблем здоров'я серед населення.

Обмежений доступ до медичної допомоги: Воєнні конфлікти можуть призводити до обмеженого доступу до медичних закладів, а також втрати медичного обладнання та ресурсів. Це робить населення більш вразливим до ефектів погіршення якості повітря.

В даному контексті важливо враховувати ускладнення, які призводять до обмеженого доступу до даних що необхідні для моделювання, аналізу та прогнозування. Але проблема чистого повітря в індустріальних регіонах була завжди актуальною і в довоєнні часи. Тому в даній роботі для аналізу були відібрані доступні відкриті дані про стан захворюваності, економіки та повітря в довоєнні часи, а саме в період з 2011 року по 2017 рік та частково відомі окремі показники по 2023р.

Метою даного дослідження є вивчення залежностей між станом економіки, якістю повітря та захворюваністю населення, а також побудова нейромережевої моделі для аналізу цих взаємозв’язків. Окрім того,

розглядаються проблеми, які ускладнюють оцінку впливу якості повітря на здоров’я населення в умовах війни.

Задачі, що дозволяють досягти мети, включають визначення ключових факторів для побудови нейромережевої моделі, а також збір і обробку відповідних даних з доступних відкритих джерел.

Практична цінність отриманих результатів зумовлена необхідністю розробки ефективних стратегій моніторингу та заходів для зниження негативного впливу воєнних дій на стан здоров’я населення та покращення системи прогнозування захворюваності.

Наукова новизна дослідження полягає у подальшому розвитку нейромережевих методів для прогнозування захворюваності та вдосконаленні стратегій мінімізації ризиків для здоров'я населення в умовах війни.

Окремий аналіз впливу розвитку альтернативних джерел енергії на загальну захворюваність населення України дозволив обґрунтувати перспективний напрямок для вирішення екологічних і соціальних проблем.

# 1 АНАЛІЗ ІСНУЮЧИХ ДОСЛІДЖЕНЬ ЕКОНОМІКИ, ВПЛИВУ ВІЙНИ НА ДОВКІЛЛЯ ТА ЗДОРОВ'Я ЛЮДЕЙ

В роботі [1] розглянуто економічну модель, що ґрунтується на моделі Солоу. Ця модель може застосовуватися для будь-якого періоду часу (включаючи як минуле, так і майбутні прогнози). Автори [1] розглянули модель в період з 2015 року до 2035 року відповідно до прогнозів моделі здоров'я на основі даних восьми індустріальних міст України, зокрема Запоріжжя. Результатом моделі є щорічний загальний ВВП України, вимірюваний в номінальних доларах США. Модель відносно точно відтворила зростання ВВП, спостережене між 2015 та 2020 роками, з прогнозованим ВВП у 2020 році на рівні 159,6 мільярда доларів США (порівняно з 155,4 мільярда доларів США, оціненими Всесвітнім банком). Прогнозоване зростання тривало експоненційно, досягаючи 278,2 мільярда доларів США до 2025 року, 500,5 мільярда доларів США до 2030 року і 940,2 мільярда доларів США до 2035 року; іншими словами, прогнозувалося подвоєння ВВП приблизно кожні 6–7 років. У порівнянні з цим робоча сила, яка вносить свій внесок у економіку країни, прогнозувалася трошки зменшитися з часом через демографічний спад. Кількість осіб, які роблять свій внесок у економічний вихід, дорівнювала 17,5 мільйонам людино-років у 2020 році, повільно зменшуватиметься до 15,0 мільйонів людино-років до 2035 року.

Вплив війни. Кілька досліджень з різним фокусом свідчать, що громадські протести або страйки також значно обмежують викиди моторизованого транспорту [2]. Однак ці виняткові події можуть створювати додаткові джерела забруднення, наприклад, через збільшення використання приватних автомобілів, спалювання і вандалізм з метою створення фізичних бар'єрів [3]. Про це вказано в загальному звіті про зниження рівнів міського забруднення [4]. Хоча існує багато публікацій, які акцентують увагу на впливі обмежень для звичайних людських активностей (наприклад, пандемії та протести), а також економічні та психологічні ефекти війни [5], досліджень війни

та її впливу на повітряне забруднення мало [6]. Деякі дослідження фокусуються на загальному розумінні впливу, який війна може мати на якість повітря та глобальну температуру з точки зору розвитку та використання зброї, а також економічного та екологічного тягаря відновлення зруйнованого війною [7]. Також цікаво оцінити екологічні ефекти атмосферних викидів після війни або, навпаки, економічної кризи [8]. Інші дослідження фокусуються на хімічний склад снарядів, ракет чи ракет, які розмелюються під час вибуху і будуть багато років в навколишньому середовищі (тобто метали тощо), постійно отруюючи середовище та здоров'я людини [9]. Є також конкретні поодинокі дослідження про обмеження Холодної війни на зусилля щодо моніторингу регіонального та глобального повітряного забруднення [10], або кількісну оцінку ризиків передчасної смертності, пов'язаної зі збільшенням рівнів PM10 через Війну в Перській затоці 1991–1992 років [11], та відповідний спад у виробництві сонячної енергії [12]. Проведені дослідження регіонального повітряного забруднення внаслідок одночасного знищення основних промислових джерел через промислові нещасні випадки чи пожежі на нафтопереробних заводах в Сербії [13]. Дуже мало досліджень вивчають зміни якості повітря під час війни, вони фокусуються на одному забруднювачі, головним чином PM. Крім того, наскільки відомо нам, немає досліджень впливу недавніх воєн на якість повітря.

Коливання рівнів якості повітря в навколишньому середовищі є серйозною проблемою, оскільки забруднення повітря є найбільшою екологічною загрозою для передчасної смертності в світі [14]. Це особливо стосується міст, де майже вся міська людська популяція дихає повітрям, що порушує рекомендації Всесвітньої організації охорони здоров'я (ВООЗ) [15]. Атмосферні забруднювачі можуть призводити до різноманітних проблем здоров'я дихальних шляхів та серцево-судинної системи [16]. Підвищені концентрації забруднення повітря особливо небезпечні для населення з високим ризиком: молодь, літні люди та люди з порушеннями здоров'ям [17].

# 2 ДАНІ ТА ПОКАЗНИКИ ДЛЯ АНАЛІЗУ ВПЛИВУ ЯКОСТІ ПОВІТРЯ

## 2.1 Огляд показників

Якість повітря можна вимірювати або за допомогою одного показника, такого як концентрація PM2,5, або за допомогою складних індексів, які враховують різні визначаючі фактори якості повітря, включаючи PM2,5, а також інші чинники. Наприклад, індекс якості повітря (AQI), який діє у Сполучених Штатах, використовує шкалу до 500 балів залежно від концентрації твердих частинок, наземного озону, чадного газу, діоксиду сірки та діоксиду азоту [15]. Рівні AQI поділяються на шість категорій: зелений (добрий), жовтий (помірний), помаранчевий (шкідливий для чутливих груп), червоний (шкідливий для здоров'я), фіолетовий (дуже шкідливий для здоров'я) та темно-бордовий (небезпечний). Хоча в різних країнах розроблялися і приймалися інші шкали, класифікація за AQI буде прийнята в даному дослідженні через його широке використання та сумісність з такими прямими показниками, як PM2,5 [18].

## 2.2 Тягар хвороб

Тягар хвороб зазвичай вимірюється в роках життя з поправкою на інвалідність (DALY). DALY – це концепція, яка об'єднує як життя, втрачене через передчасну смерть, так і життя, прожите з інвалідністю або іншим станом, який серйозно впливає на здатність брати участь і вносити внесок в життя суспільства. Коротко кажучи, DALY визначається як сума втрачених років життя (YLL) та років життя, прожитих з інвалідністю (YLD). YLL розраховується як різниця фактичного часу смерті та максимально можливого терміну життя людини на цьому віці в даній країні. Останній визначається з використанням оцінок на рівні країни з прогнозів World Population Prospect на 2050 рік. YLD

визначається шляхом множення тривалості життя з певним станом здоров'я на вагу інвалідності, яка відображає важкість захворювання.

Показник DALY використовується поряд із кількістю передчасних смертей та показниками конкретних захворювань (захворюваність, поширеність) як основні показники тягара для здоров'я. Крім того, компоненти DALY (YLL і YLD) будуть використовуватися в якості вхідних параметрів економічної моделі, де вони служитимуть показниками втрати або потенційного внеску в робочу силу [6].

## 2.3 Економічна модель

ВВП розраховується як добуток TFP, фізичного та людського капіталу. TFP (Загальна продуктивність факторів виробництва) (до 2017 року), рівень збережень та рівень амортизації адаптовані зі світової таблиці Пенна (це міжнародна база даних, яка надає інформацію про різноманітні економічні показники та агрегати для більшості країн світу).

Людський капітал розраховується у кілька етапів. Спочатку дані з програмного інструменту DemProj (Avenir Health/Spectrum package) використовуються для оцінки населення в кожній віковій та статевій групі (5-річні інтервали) кожного року [18]. Населення в кожній групі помножується на коефіцієнт участі, адаптований з бази даних Міжнародної організації праці, що визначає частку осіб, які роблять свій внесок у економічне виробництво країни[19]. З отриманої активної робочої сили в кожній віковій та статевій групі віднімаються YLD (взяті з GBD; всі причини) [20]. Нарешті, внесок залишеної робочої сили помножується на індекс людського капіталу (світова таблиця Пенна) [21]. Економічна модель запускалася в тих самих чотирьох сценаріях, що й модель прогнозу здоров'я на національному рівні, представленої в попередньому розділі: базовий сценарій (якість повітря залишається постійною)

або поступове зниження концентрації PM2.5 до 15, 10 або 5 µg/m3 до 2030 року. Параметризація моделі представлена на Рисунку 2.3.1. Окрім параметрів, показаних вище, дані щодо чисельності населення та віково-статевої структури були адаптовані з інструменту DemProj [18], а рівні економічної участі за віком і статтю з бази даних ILOSTAT [19]. Також враховано показники валової продуктивності праці та загальної продуктивності факторів.

| Indicator | Value | Source |
|---|---|---|
| Initial values in 2015 | | |
| Total factor productivity | 1.707 | Penn World Table (*21*) |
| Physical capital stock | 7205.3 billion USD | Penn World Table (*21*) |
| Human capital index | 3.2554 | Penn World Table (*21*) |
| Constants | | |
| Depreciation rate | 2% | Assumption |
| Savings rate | 40% | Assumption |
| Growth of TFP | 0.12 | Fitted |
| Elasticity relative to physical capital | 0.5 | Assumption |
| Fitting coefficient | 2.6 | Fitted |

Рисунок 2.1 – Вхідні параметри для економічної моделі [18]

# 3 ПІДГОТОВКА ДАНИХ ДО АНАЛІЗУ

Для проведення аналізу були використані дані з відкритих джерел [22–23], що охоплюють низку соціально-економічних та екологічних показників. Зокрема, бюджет ОЗ – це сума коштів, офіційно виділена на фінансування системи охорони здоров’я в межах державного бюджету. Обсяг викидів забруднюючих речовин представлений в умовних одиницях і відображає рівень забруднення довкілля, що може впливати на стан здоров’я населення. Валовий внутрішній продукт (ВВП) – це загальна вартість усіх товарів і послуг, вироблених у країні протягом року, що служить показником економічного розвитку. Рівень використання відновлюваних джерел енергії визначає частку "зеленої" енергії в загальній структурі енергоспоживання країни і є важливим з погляду сталого розвитку та екологічної безпеки.

Таблиця 3.1 – Вхідні дані для подальшої обробки

| Рік | Бюджет ОЗ | Викиди забруднюючих речовин | ВВП | Використання відновлюваних джерел |
|---|---|---|---|---|
| 2011 | 7,5 | 4374,6 | 29980 | 2514 |
| 2012 | 7 | 4335,3 | 32480 | 2476 |
| 2013 | 10 | 4295,1 | 33965 | 3166 |
| 2014 | 8,5 | 3350 | 36904 | 2797 |
| 2015 | 11,4 | 2857,4 | 46413 | 2700 |
| 2016 | 12,1 | 3078,1 | 55899 | 3616 |
| 2017 | 16,4 | 2584,9 | 70170 | 3907 |
| 2018 | 26,5 | 2508,3 | 84228 | 4303 |
| 2019 | 39,5 | 2459,5 | 94633 | 4335 |
| 2020 | 116,4 | 2238,6 | 101138 | 5687 |

Валютна нестабільність гривні може впливати на результати, тому її слід брати до уваги. Мною було прийнято рішення нормувати ці значення, а саме Бюджет ОЗ та ВВП перевести в долари. Для цього використовуючи парсинг (Додаток А) був взятий середній курс долара за кожен рік [24]. Далі поділив значення вибраних комірок таблички на відповідний курс долара.

Дані, які стосуються захворюваності України були отримані з офіційного запиту до «Центра медичної статистики Міністерства охорони здоров'я України» [25]. У відповідь був отриманий файл зі скан-копіями звітів, який слід було переписати в табличний вид. Для цього було використано сайт для конвертації файлів [26], після чого файл перевірявся та були заповнені пропуски і неточності через погану якість скан-копій. Після цього файл треба привести до вигляду, придатного для аналізу, для цього було використано програму для створення більш зручної таблиці (Додаток Б).

Для детальної перевірки був написаний скрипт, який виявляє в якій з комірок міг бути пропущений чи доданий зайвий знак до числа (Додаток В).

Далі треба нормувати ці дані, адже в 2014 була окупована частина територій України, що спричинило суттєву зміну кількості населення, на основі якої надаються статистичні дані. Для цього населення України в період 2011–2020 років [27] було поділено на кількість захворюваних відповідного року (Додаток Г).

Тепер можна розглядати самі значення кількості захворюваних і чи є між ними якась залежність. На мою думку, варто оцінити, наскільки графіки різних категорій є лінійними або демонструють суттєві відхилення від лінійної залежності. Для цього був написаний скрипт (Додаток Д). Завдяки цьому коду були виявлені топ аномальних (рис. 3.1) та топ лінійних (рис. 3.2) залежностей.

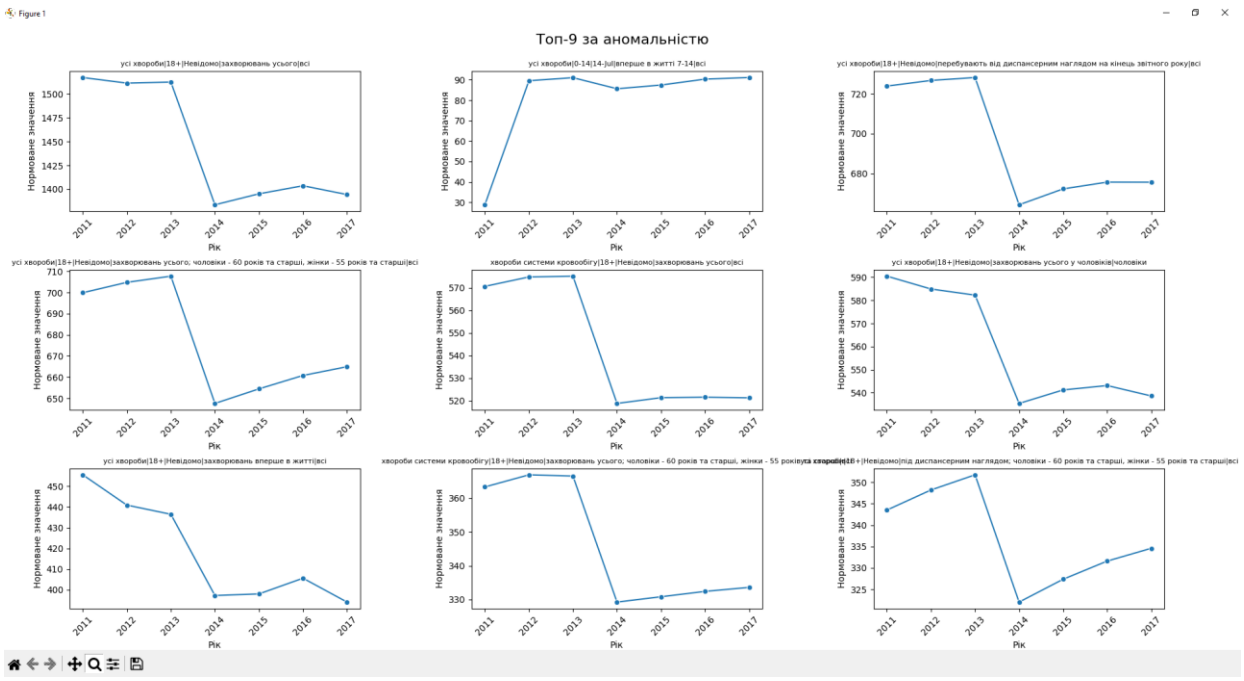



Рисунок 3.1 – Топ-9 аномальних залежностей

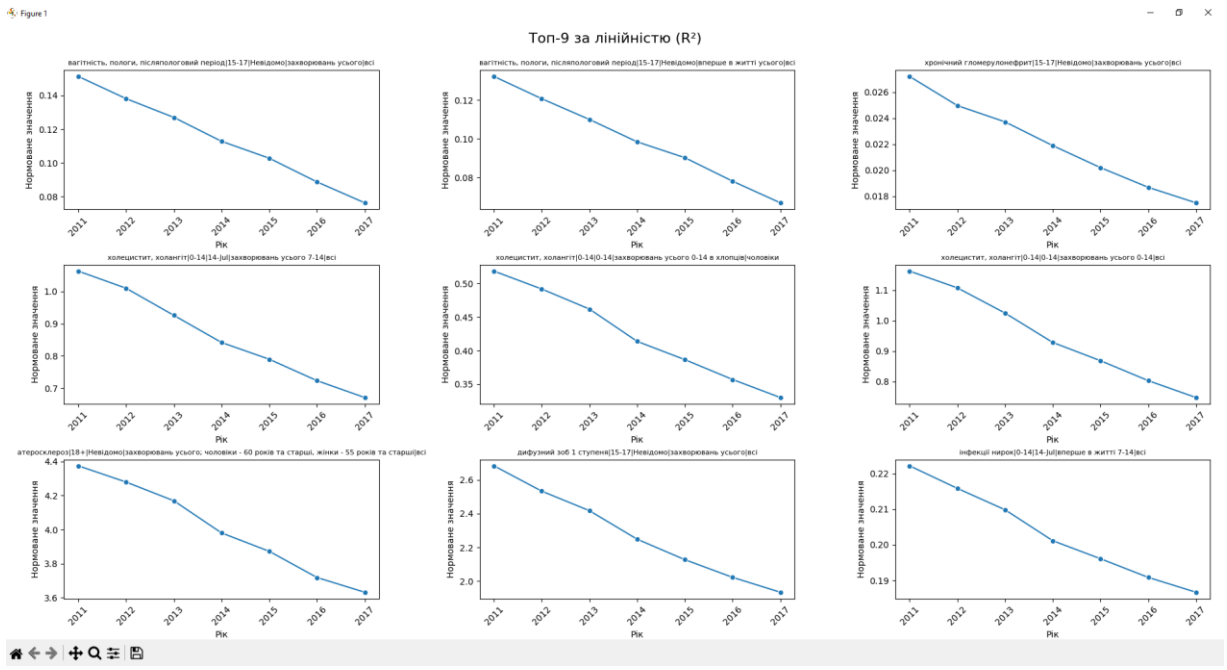



Рисунок 3.2 – Топ-9 лінійних залежностей

Для отриманих та збережених даних, вибираємо що саме будемо аналізувати, наприклад «холецистит, холангіт|0-14|0-14|захворювань усього 0-14|всі». Для цього в файлі «усі_захворювання_tidied_нормовані» робимо фільтрацію по відповідних стовпцях (рис. 3.3), беремо дані стовпчика «нормовані» та додаємо їх в таблицю для аналізу (табл. 3.2).

| Хвороб | Рік | Вікова | Вікова | Катего | Стать | value | naseler | Нормо | не |
|---|---|---|---|---|---|---|---|---|---|
| холецисті | 2011 | 0-14 | 0-14 | захворюв | всі | 53233 | 45778.5 | 1.162838 | |
| холецисті | 2012 | 0-14 | 0-14 | захворюв | всі | 50558 | 45633.6 | 1.107912 | |
| холецисті | 2013 | 0-14 | 0-14 | захворюв | всі | 46692 | 45553 | 1.025004 | |
| холецисті | 2014 | 0-14 | 0-14 | захворюв | всі | 40004 | 43087.7 | 0.928432 | |
| холецисті | 2015 | 0-14 | 0-14 | захворюв | всі | 37281 | 42928.9 | 0.868436 | |
| холецисті | 2016 | 0-14 | 0-14 | захворюв | всі | 34334 | 42760.5 | 0.802937 | |
| холецисті | 2017 | 0-14 | 0-14 | захворюв | всі | 31829 | 42584.5 | 0.747432 | |

Рисунок 3.3 – Результат фільтрації даних

Таблиця 3.2 – Дані для подальшого аналізу

| Рік | Бюджет ОЗ | Викиди забруднюючих речовин | ВВП | Використання відновлюваних джерел | холецистит, холангіт\|0-14\|захворювань усього 0-14\|всі |
|---|---|---|---|---|---|
| 2011 | 0,9475 | 4374,6 | 3787,6 | 2514 | 1,162838 |
| 2012 | 0,876 | 4335,3 | 4064,6 | 2476 | 1,107912 |
| 2013 | 1,2511 | 4295,1 | 4249,3 | 3166 | 1,025004 |
| 2014 | 0,7111 | 3350,0 | 3087,3 | 2797 | 0,928432 |
| 2015 | 0,5211 | 2857,4 | 2121,4 | 2700 | 0,868436 |
| 2016 | 0,4729 | 3078,1 | 2184,6 | 3616 | 0,802937 |
| 2017 | 0,6168 | 2584,9 | 2638,9 | 3907 | 0,747432 |
| 2018 | 0,9741 | 2508,3 | 3096,2 | 4303 | |
| 2019 | 1,5307 | 2459,5 | 3667,2 | 4335 | |
| 2020 | 4,3118 | 2238,6 | 3746,4 | 5687 | |

# 4 ПОБУДОВА НЕЙРОМЕРЕЖІ

Для аналізу даних будемо використовувати нейромережу, а саме бібліотеку Keras [28] — це високорівнева бібліотека для машинного навчання на Python, створена для швидкої і зручної побудови нейронних мереж. Вона надає простий і зрозумілий інтерфейс для створення моделей глибокого навчання, таких як згорткові та рекурентні нейронні мережі, і працює на основі фреймворка TensorFlow. Завдяки своїй простоті та потужності, Keras дозволяє швидко навчати та оцінювати моделі, що робить її популярним вибором серед дослідників і розробників у галузі штучного інтелекту.

Для реалізації машинного навчання нам знадобляться наступні бібліотеки та модулі:

```
import pandas as pd
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import Input
from tensorflow.keras.optimizers import Adam
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
import numpy as np
```

Для першого випробування побудуємо нейромережу з двома шарами, де на першому шарі будуть 16 нейронів, а на другому 12 (Додаток Е), Навчання продоводилось протягом 150 епох. Результат (рис. 4.1) можна визначити як точний, середнє квадратичне відхилення на тренувальних даних досить мале(округлення до сотих):

Mean Squared Error on Test Data: 0,00

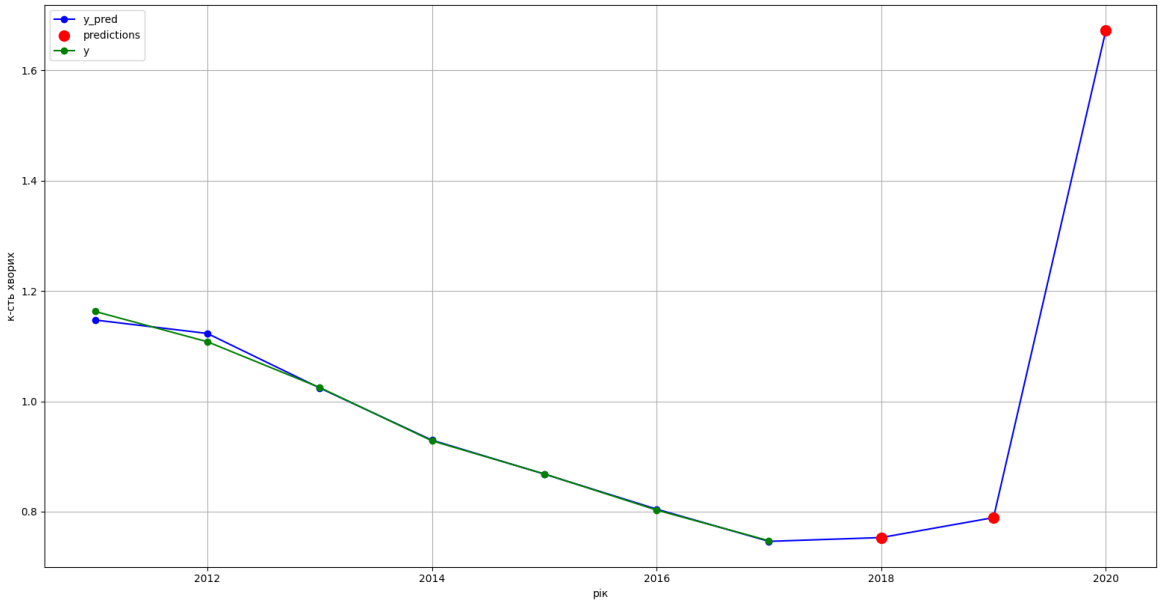



Рисунок 4.1 – Результат тестування нейронної мережі, що складається з двох шарів (16,12) та 150 епох

Такий результат свідчить про достатньо точну модель. На графіку видно суттєве зростання значень у 2020 році, що, ймовірно, пов'язано зі спалахом Covid-19. Попри те, що пандемія могла вплинути лише на окремі групи захворювань, її наслідки ускладнення та загальне навантаження на систему охорони здоров'я мали вплив майже на всі категорії. У цьому контексті отримана картина виглядає логічною.

Також варто перевірити як будуть поводити себе ці ж дані, але при трохи інших налаштуваннях нейромережі, наприклад для 500 епох:

model.fit(x1, y, epochs=500, batch_size=16)

Результат тестування після більш тривалого навчання очікувано кращий (рис. 4.2). Тепер за графіком не видно розбіжностей між фактичними даними (y), та даними отриманими нейромережею (y_pred). Тепер ми бачимо невеликий скачок у 2019 та великий (але не такий, як у минулому графіку) у 2020. Тим не менш, необхідне подальше тестування мережі, щоб впевнитись, що така висока точність на тренувальних даних не призвела до перенавчання.

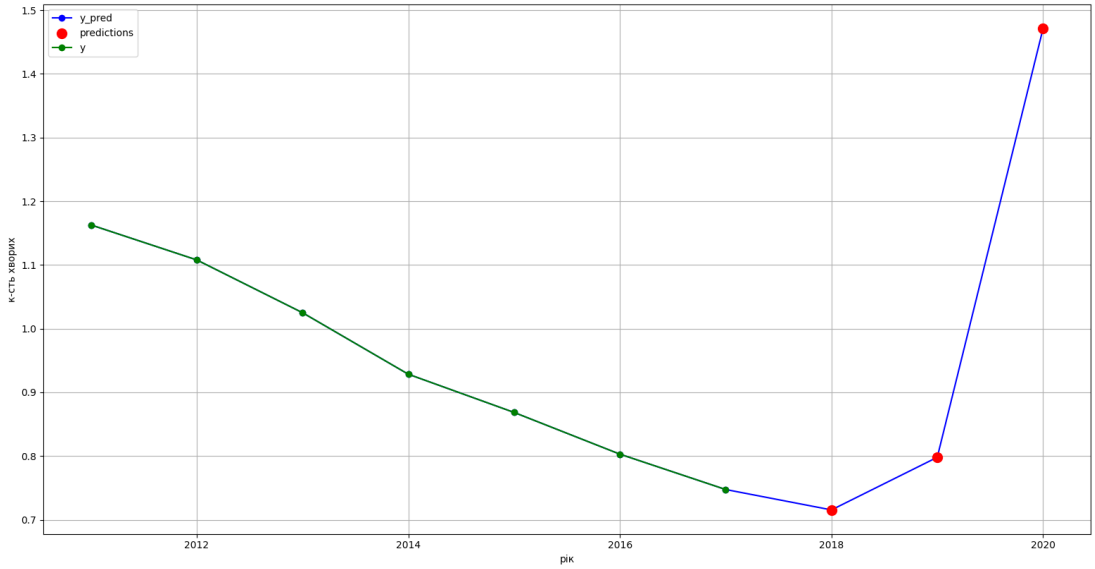



Рисунок 4.2 – Результат тестування нейронної мережі, що складається з двох шарів (16,12) та 500 епох

Розглянемо тепер тришарову нейромережу з шарами 16, 12, 8 та 150 епох, це можна зробити трохи змінивши код:

```
model = Sequential()
model.add(Input(shape=(4,)))
model.add(Dense(16, activation='relu'))
model.add(Dense(12, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1))
learning_rate = 0.05
optimizer = Adam(learning_rate=learning_rate)
model.compile(loss='mean_squared_error', optimizer=optimizer, metrics=['mae'])
model.fit(x1, y, epochs=150, batch_size=16)
```

Результат який можна побачити (рис. 4.3) приблизно такий самий, як був в першому випадку, але має навіть ще більший скачок.

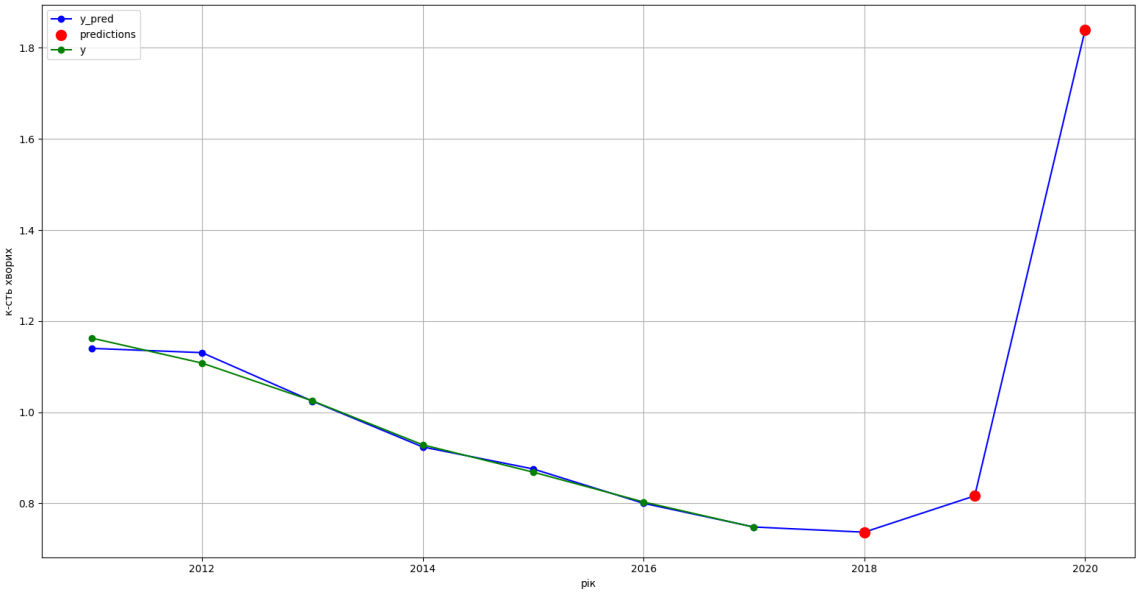



Рисунок 4.3 – Результат тестування нейронної мережі, що складається з трьох шарів (16,12,8) та 150 епох

Тепер слід для тих самих нейронних шарів збільшити кількість епох: model.fit(x1, y, epochs=500, batch_size=16)

Результати виконання такого коду досить сильно відрізняються одне від одного (рис. 4.4–4.5). На першому графіку видно різке падіння, а згодом — стрімке зростання. Другий графік порівняно з фактичними даними має недостатню точність і демонструє неприродну поведінку у ділянках, де відсутні реальні значення.

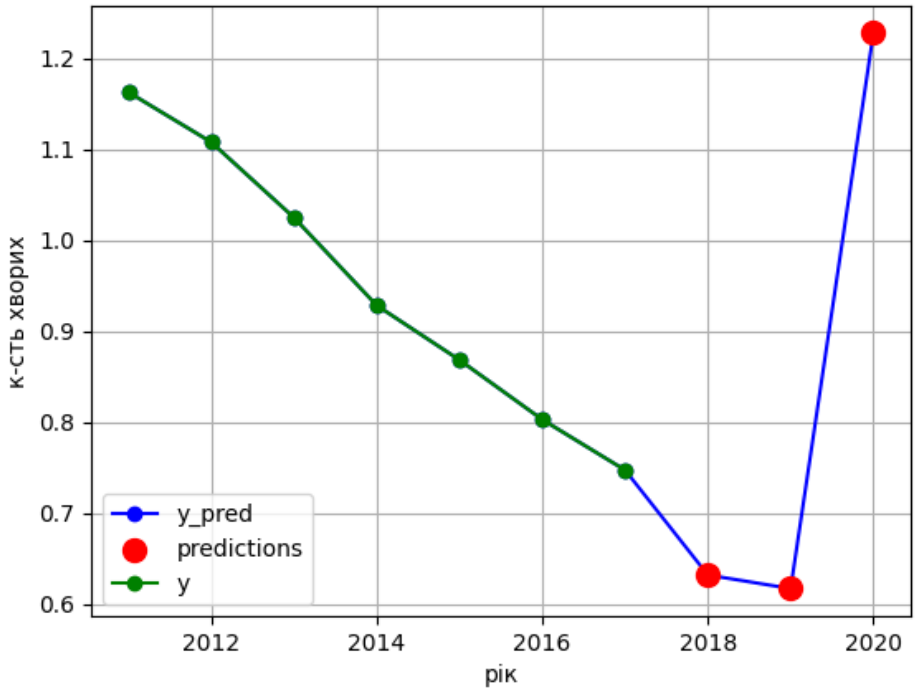



Рисунок 4.4 – Результат тестування нейронної мережі, що складається з трьох шарів (16,12,8) та 500 епох

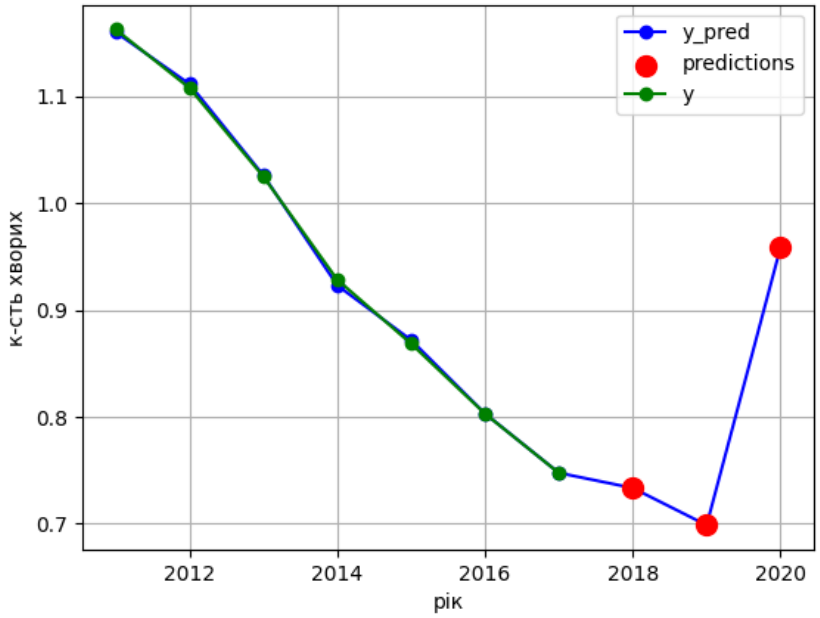



Рисунок 4.5 – Результат тестування нейронної мережі, що складається з трьох шарів (16,12,8) та 500 епох

Наступним кроком доцільно розглянути тришарову нейронну мережу з 20, 16 та 12 нейронами в кожному відповідному шарі. Фрагмент коду, що реалізує це, має такий вигляд:

```
model = Sequential()
model.add(Input(shape=(4,)))
model.add(Dense(20, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(12, activation='relu'))
model.add(Dense(1))
```

Спочатку встановимо кількість епох рівною 150, як і в інших експериментах. Отриманий результат (рис. 4.6) загалом схожий на багато попередніх варіантів, проте перші декілька значень дещо не збігаються, хоча середньоквадратичне відхилення, округлене до сотих, також дорівнює нулю.

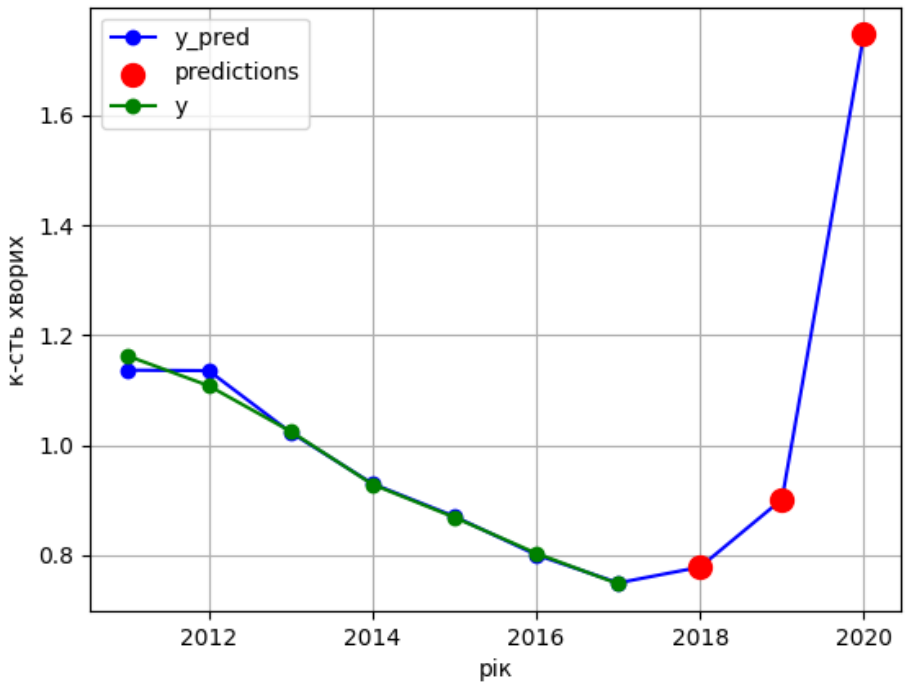



Рисунок 4.6 – Результат тестування нейронної мережі, що складається з трьох шарів (20,16,12) та 150 епох

Далі розглянемо таку саму модель шарів, але на 500 епох. При такій нейромережі результати (рис. 4.7–4.8) виходять знов не однозначні. Другий варіант показує загальну тенденцію всіх інших варіантів, а перший виглядає дещо дивно.

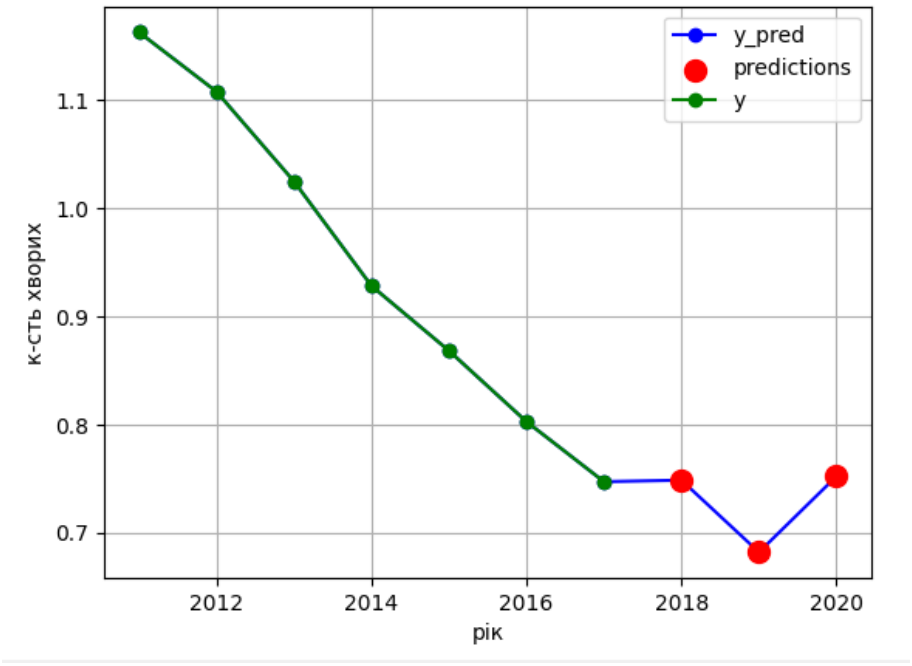



Рисунок 4.7 – Результат тестування нейронної мережі, що складається з трьох шарів (20,16,12) та 500 епох

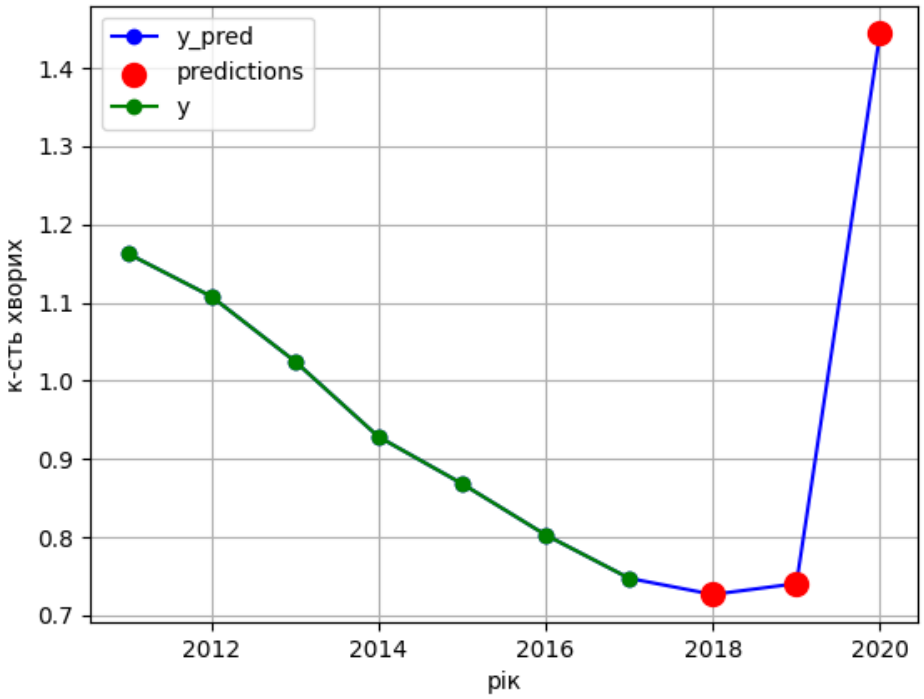
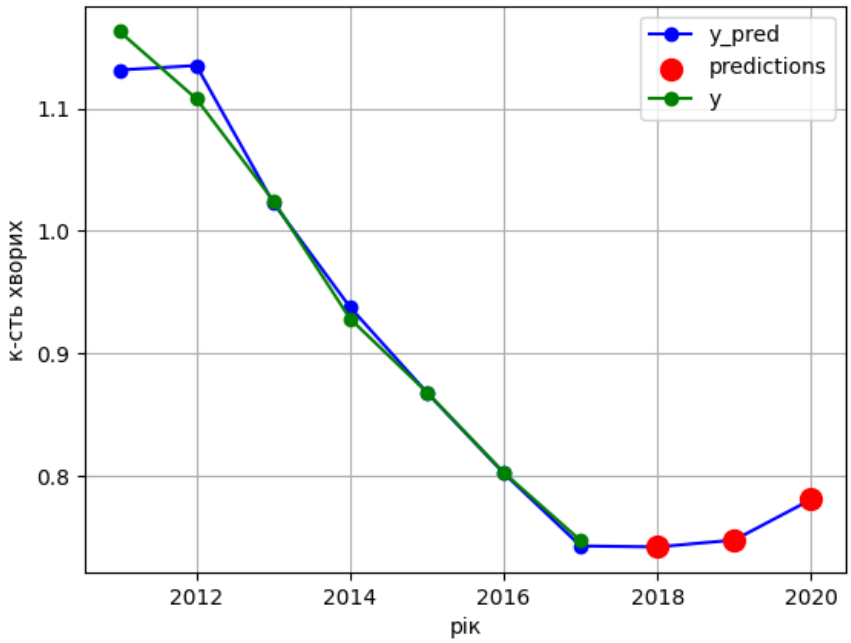



Рисунок 4.8 – Результат тестування нейронної мережі, що складається з трьох шарів (20,16,12) та 500 епох

Для завершального експерименту перевіримо вплив ще одного додаткового шару на модель. При цьому змінюємо кількість нейронів у шарах на 32, 16, 8 та 4 відповідно. Відповідний фрагмент коду матиме такий вигляд:

```
model.add(Input(shape=(4,)))
model.add(Dense(32, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(4, activation='relu'))
model.add(Dense(1))
```

Кількість епох установимо стандартно — 150. Результат (рис. 4.9–4.10) вийшов подібним до попередніх експериментів: один графік менш точний і не демонструє різкого стрибка у 2020 році, тоді як інший є точнішим і фіксує чітку зміну результату саме у 2020 році.




Рисинок 4.9 – Результат тестування нейронної мережі, що складається з чотирьох шарів(32, 16, 8, 4) та 150 епох

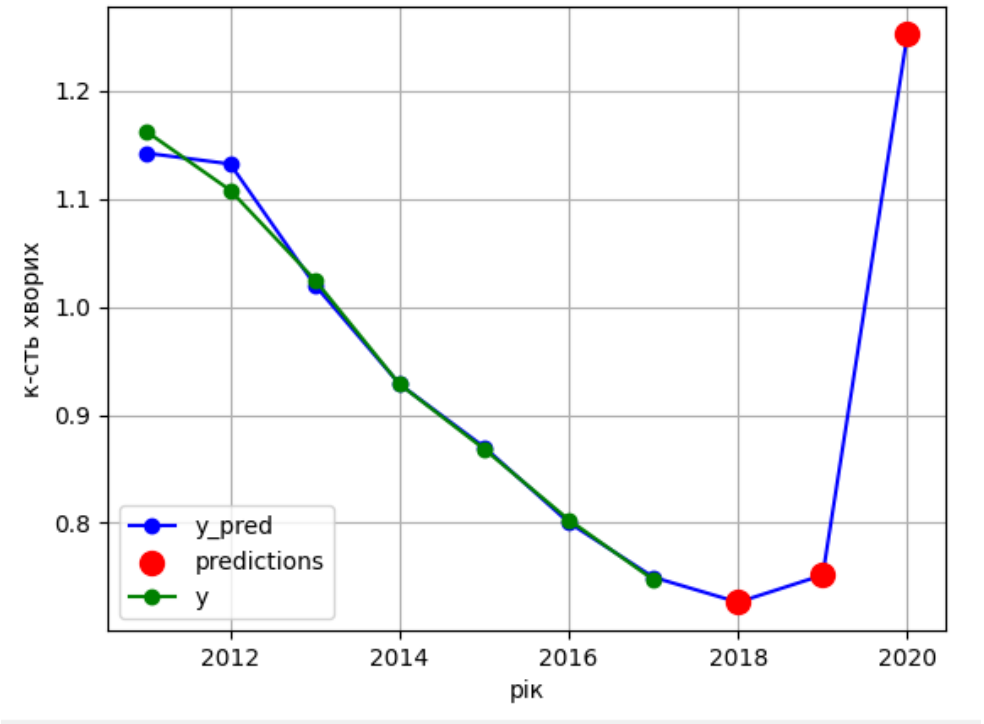



Рисунок 4.10 – Результат тестування нейронної мережі, що складається з чотирьох шарів(32, 16, 8, 4) та 150 епох

Тепер пропоную розглянути дещо іншу модель залежності: взяти лише «Викиди забруднюючих речовин», «Використання відновлювальних джерел» та кількість захворювань (табл. 4.1). Такий підхід виключає вплив фактора «Бюджету ОЗ» на аномальні дані за 2020 рік і дозволяє оцінити, чи залежить кількість захворювань винятково від якості повітря. Розглянемо кілька моделей. Першу з них побудовано як двошарову мережу (16 і 12 нейронів) із 150 епохами навчання. Її результати (рис. 4.11) виглядають досить обнадійливо: у період 2011–2017 років забруднення повітря зменшувалося, використання відновлювальних джерел зростало, а захворюваність падала, тому показники для 2018–2020 років, коли тенденції якості повітря залишаються стабільними, цілком узгоджуються з очікуваннями.

Таблиця 4.1 – дані до нової моделі

| Рік | Викиди забруднюючих речовин | Використання відновлюваних джерел | холецистит, холангіт\|0-14\|захворювань усього 0-14\|всі |
|---|---|---|---|
| 2011 | 4374,6 | 2514 | 1,06321 |
| 2012 | 4335,3 | 2476 | 1,01018 |
| 2013 | 4295,1 | 3166 | 0,92606 |
| 2014 | 3350 | 2797 | 0,84179 |
| 2015 | 2857,4 | 2700 | 0,78972 |
| 2016 | 3078,1 | 3616 | 0,72331 |
| 2017 | 2584,9 | 3907 | 0,67097 |
| 2018 | 2508,3 | 4303 | |
| 2019 | 2459,5 | 4335 | |
| 2020 | 2238,6 | 5687 | |

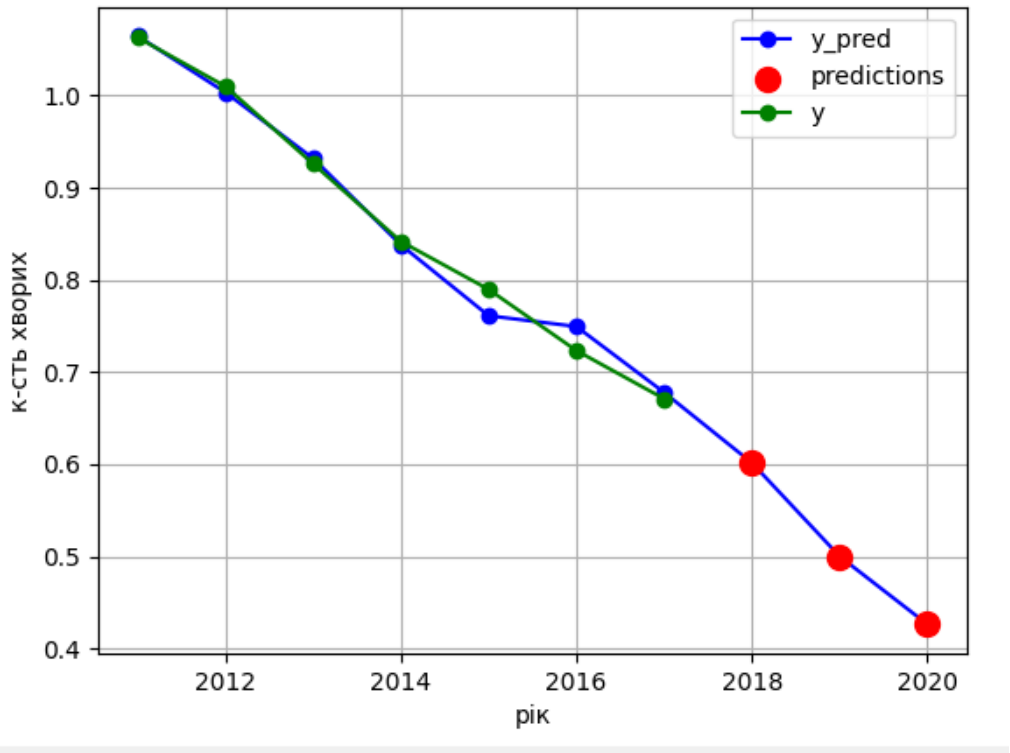



Рисунок 4.11 –Результат виконання двох шарів (16, 12), 150 епох від двох змінних

Для перевірки лінійності цієї моделі застосуємо наступний алгоритм:

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
X_lin = np.arange(2011, 2021).reshape(-1, 1)
y_lin = combined_array.reshape(-1, 1)
model_lin = LinearRegression()
model_lin.fit(X_lin, y_lin)
y_pred_lin = model_lin.predict(X_lin)
r2 = r2_score(y_lin, y_pred_lin)
print(f"Коефіцієнт детермінації R²: {r2:.4f}")
```

Коефіцієнт детермінації – це метрика, яка показує, наскільки добре лінійна модель пояснює залежність змінної. В результаті виконання цього коду був отриманий результат «Коефіцієнт детермінації R²: 0.9907», що говорить про досить високий лінійний зв’язок.

Тепер розглянемо лінійні, але не нормовані дані. Для цього в скрипті для знаходження лінійних та аномальних залежностей(Додаток Д) замінимо в усьому коді «нормовані значення» на «value» для аналізу початкових значень. Для подальшого аналізу я вибрав дані з категорії «дифузний зоб 2–3 ступеня|15-17|Невідомо|захворювань усього|всі» зі значенням коефіцієнту детермінації 0.9954, що говорить про досить велику лінійну залежність. Для початку розглянемо цю залежність від усіх чотирьох параметрів (табл. 4.2).

Таблиця 4.2 – дані для подальшого аналізу

| Рік | Бюджет ОЗ | Викиди забруднюючих речовин | ВВП | Використання відновлюваних джерел | дифузний зоб 2-3 ступеня\|15-17 \|захворювань усього |
|---|---|---|---|---|---|
| 2011 | 0,95 | 4374,6 | 3787,55 | 2514 | 12990 |
| 2012 | 0,88 | 4335,3 | 4064,57 | 2476 | 12433 |
| 2013 | 1,25 | 4295,1 | 4249,34 | 3166 | 11618 |
| 2014 | 0,71 | 3350 | 3087,35 | 2797 | 10743 |
| 2015 | 0,52 | 2857,4 | 2121,42 | 2700 | 9832 |
| 2016 | 0,47 | 3078,1 | 2184,64 | 3616 | 8823 |
| 2017 | 0,62 | 2584,9 | 2638,92 | 3907 | 8141 |
| 2018 | 0,97 | 2508,3 | 3096,23 | 4303 | |
| 2019 | 1,53 | 2459,5 | 3667,25 | 4335 | |
| 2020 | 4,31 | 2238,6 | 3746,45 | 5687 | |

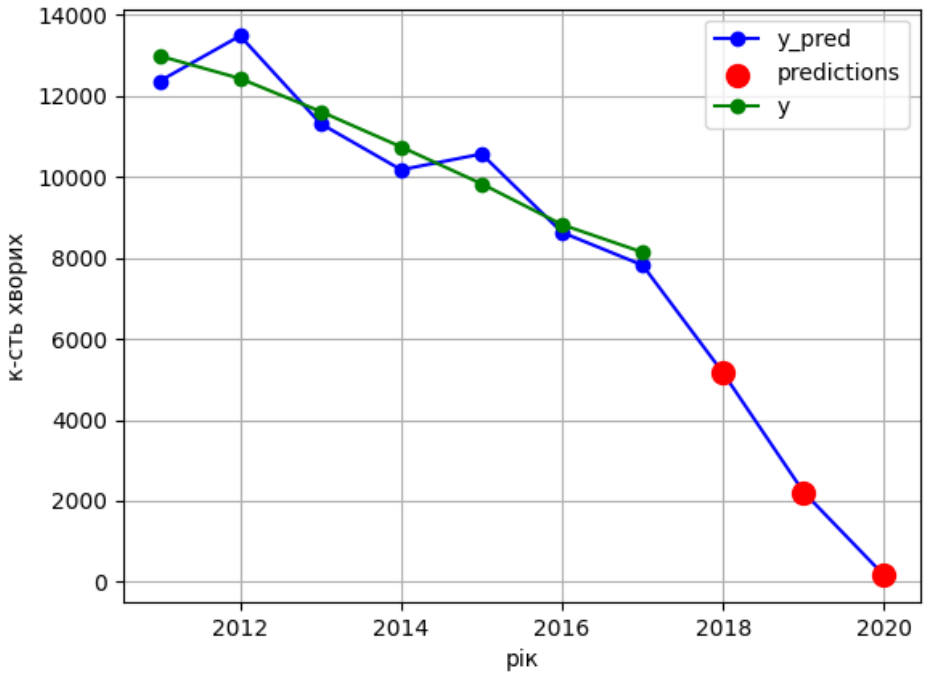



Рисунок 4.12 – Результат тестування нейронної мережі, що складається з двох шарів(16, 12) та 150 епох

Для початку візьмемо два шари з 16 та 12 нейронами відповідно та 150 епох. Результат виконання такої залежності (рис. 4.12) виглядає досить неточним, і в цьому випадку краще розглянути покращену версію нейронної мережі.

Для цього додамо третій шар та змінимо кількість нейронів, тепер буде 16, 8 та 4 на трьох шарах відповідно, але результат (рис. 4.13) дає приблизно ті самі дані, що нам не дуже підходить. Для наступного кроку спробуємо збільшити кількість епох до 500. Результат, який можна побачити, дає приблизно ті самі дані, але більш точні (рис. 4.14).

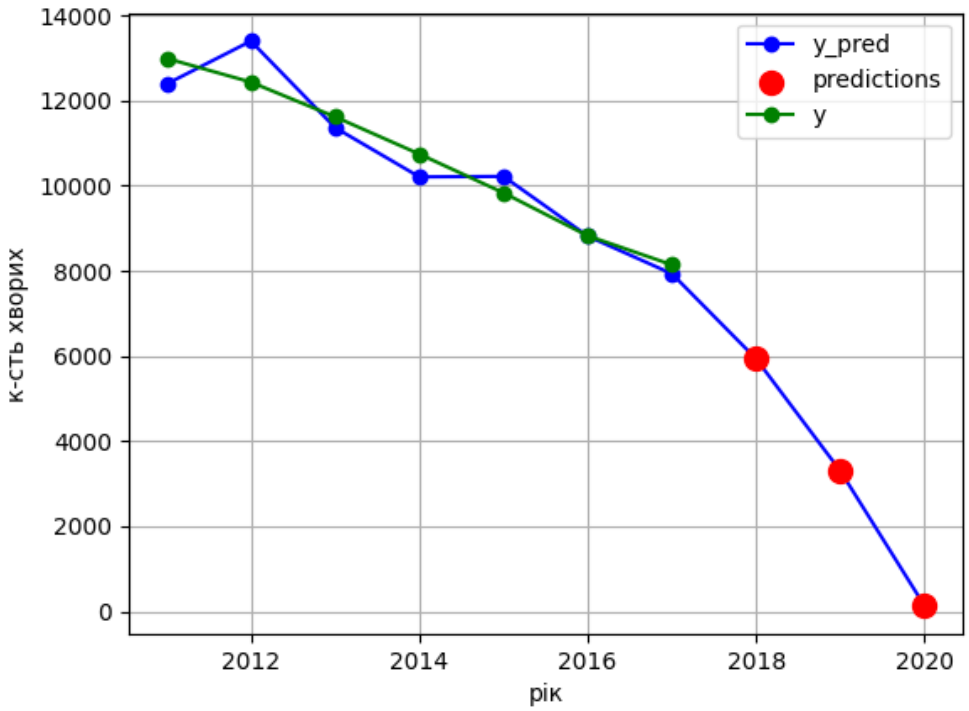



Рисунок 4.13 – Результат тестування нейронної мережі, що складається з трьох шарів (16, 8, 4) та 150 епох

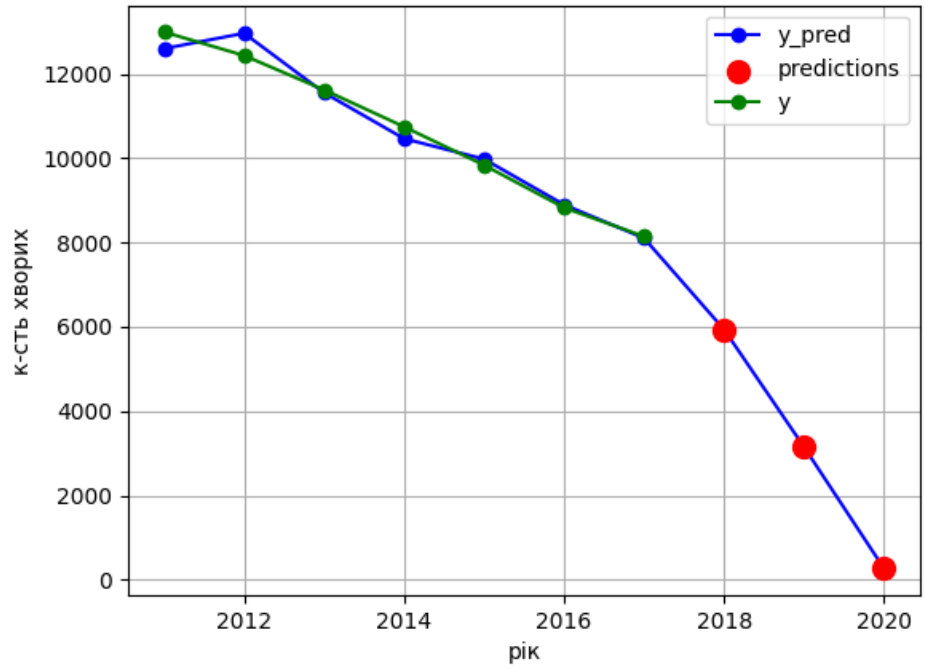



Рисунок 4.14 – Результат тестування нейронної мережі, що складається з трьох шарів (16, 8, 4) та 500 епох

Якщо виключити ВВП і розглядати лише три змінні, то при таких налаштуваннях нейронної мережі отримуємо наступні результати: двошарова з 16 і 8 нейронами та 150 епох (рис. 4.18); двошарова з 16 і 8 нейронами та 500 епох (рис. 4.19); тришарова з 16, 12 і 8 нейронами та 150 епох (рис. 4.20); тришарова з 16, 12 і 8 нейронами та 500 епох (рис. 4.21); чотиришарова з 32, 16, 12 і 8 нейронами та 500 епох (рис. 4.22). За результатами видно, що вони зовсім неточні й не мають лінійного характеру, за винятком тришарової та чотиришарової моделей на 500 епох. Це свідчить про низьку якість прогнозів. Крім того, в більшості випадків прогноз на 2020 рік близький до нуля, що також не відображає реалістичності даних. Вважаю, що в обох експериментах так сталося через відсутність нормалізації даних про захворюваність. Отже, наступну мережу слід розглядати без ВВП, але з попередньою нормалізацією показників захворюваності.

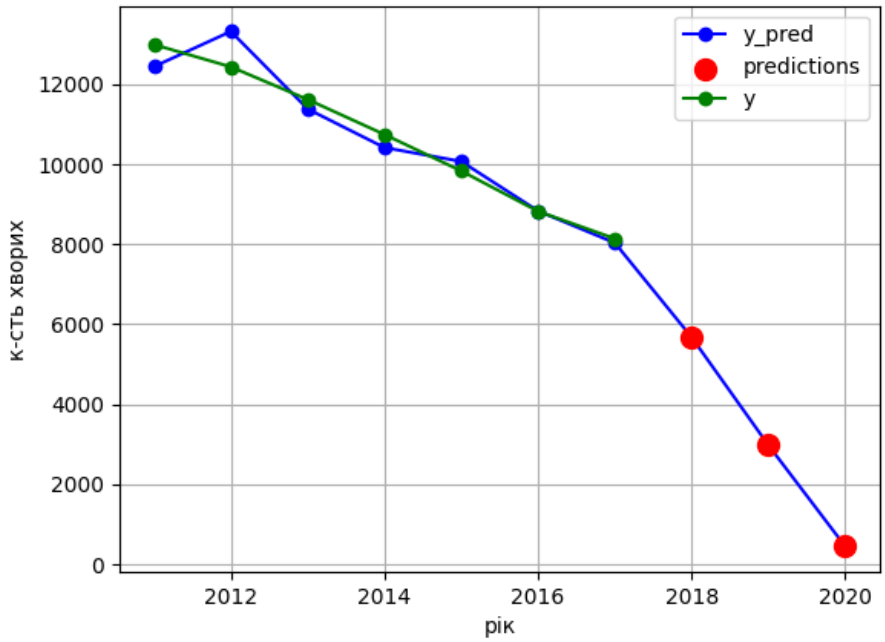



Рисунок 4.15 – Результат тестування нейронної мережі, що складається з чотирьох шарів(32, 16, 8, 4) та 150 епох

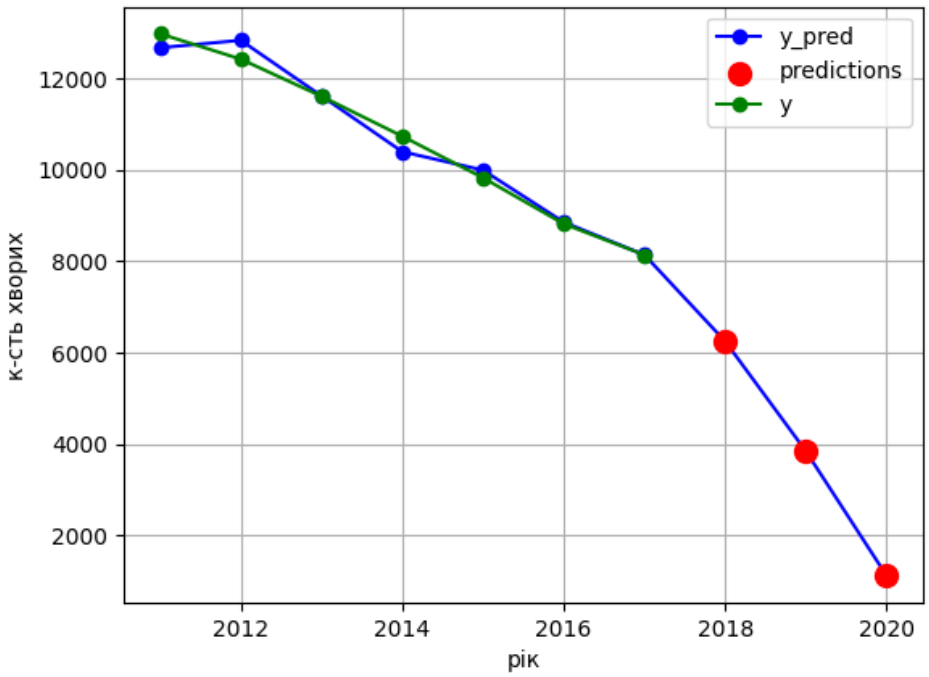



Рисунок 4.16 – Результат тестування нейронної мережі, що складається з чотирьох шарів(32, 16, 8, 4) та 500 епох

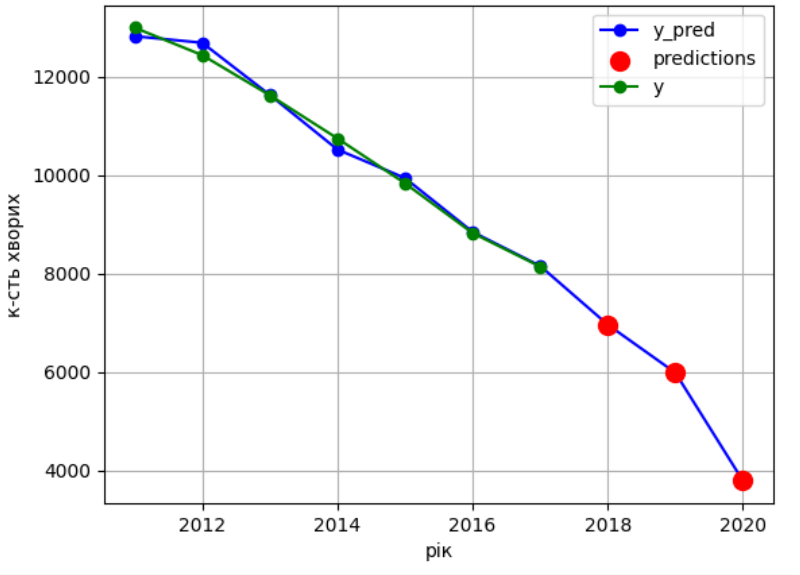



Рисунок 4.17 – Результат тестування нейронної мережі, що складається з чотирьох шарів(32, 16, 8, 4) та 1000 епох

Якщо прибрати ВВП та спробувати розглядати від трьох змінних то при наступних показниках нейронної мережі виходять такі результати: двошарова з 16, 8 нейронами та 150 епох (рис. 4.18); двошарова з 16, 8 нейронами та 500 епох (рис. 4.19); тришарова з 16, 12, 8 нейронами та 150 епох (рис. 4.20); тришарова з 16, 12, 8 нейронами та 500 епох (рис. 4.21); чотирьох шарова з 32, 16, 12, 8 нейронами там 500 епох (рис. 4.22). По результатам можна сказати, що вони зовсім не точні, навіть не лінійні окрім трьох шарової та чотирьох шарової на 500 епох. Це говорить про не дуже гарні результати. Так само можна сказати, що в більшості випадків прогнозуючий результат за 2020 рік прагне до нуля, що теж не говорить нам про реалістичність цих даних. Я вважаю, що як в цьому випадку, так і в минулому, це відбувається через те, що в цих системах дані про захворюваність не були пронормовані. Тому, на мою думку, наступну мережу треба розглянути без ВВП, але з нормуванням захворюваних.

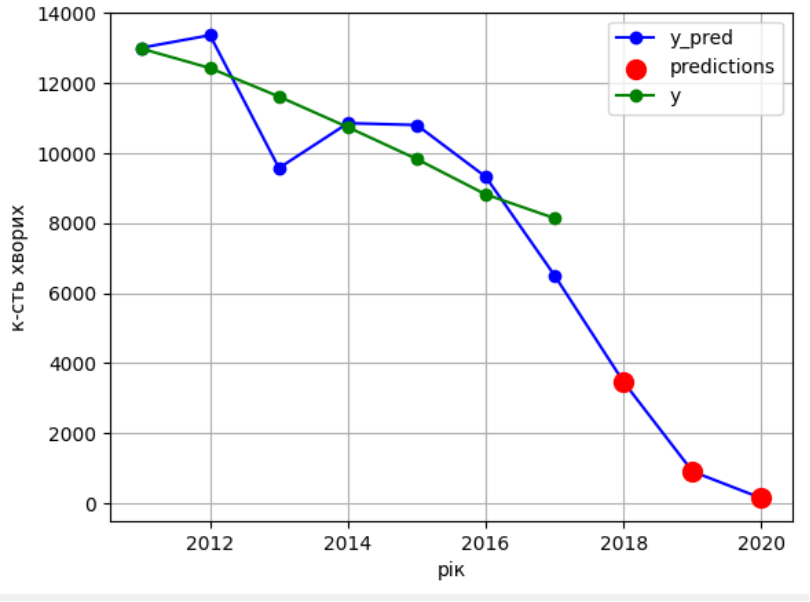



Рисунок 4.18 – Результат тестування нейронної мережі, що складається з двох шарів(16, 8) та 150 епох

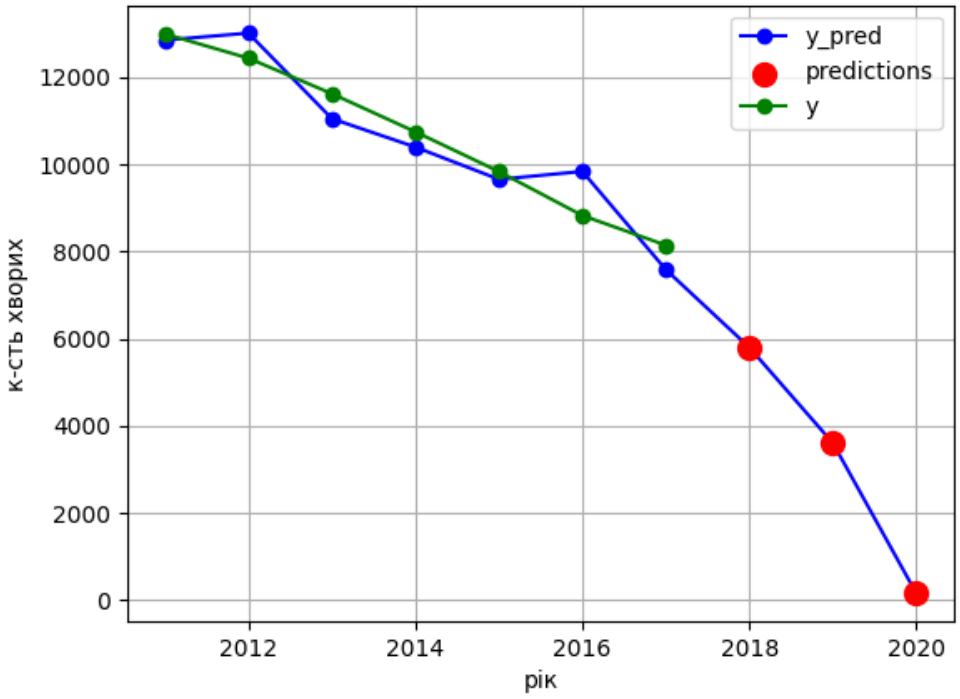



Рисунок 4.19 – Результат тестування нейронної мережі, що складається з двох шарів(16, 8) та 500 епох

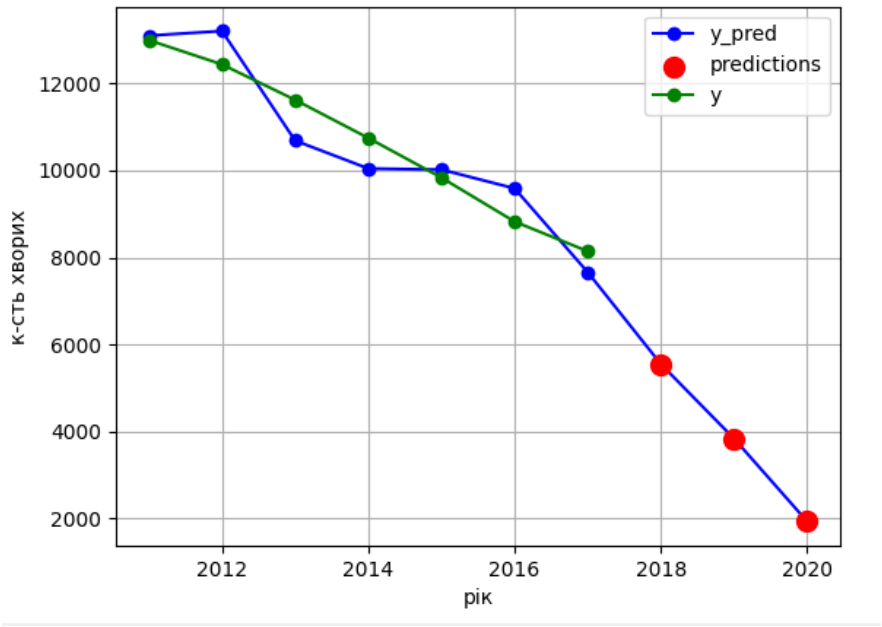



Рисунок 4.20 – Результат тестування нейронної мережі, що складається з трьох шарів(16, 12, 8) та 150 епох

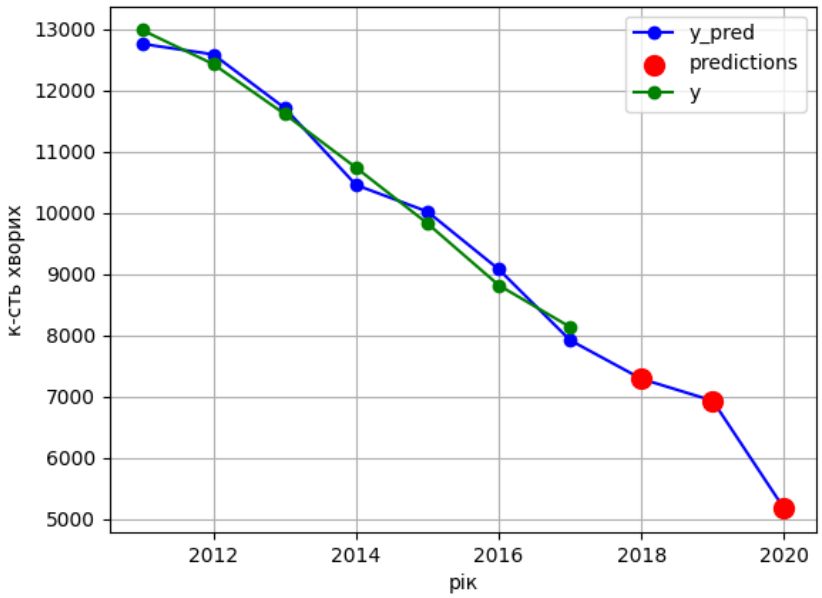



Рисунок 4.21 – Результат тестування нейронної мережі, що складається з трьох шарів(16, 12, 8) та 500 епох

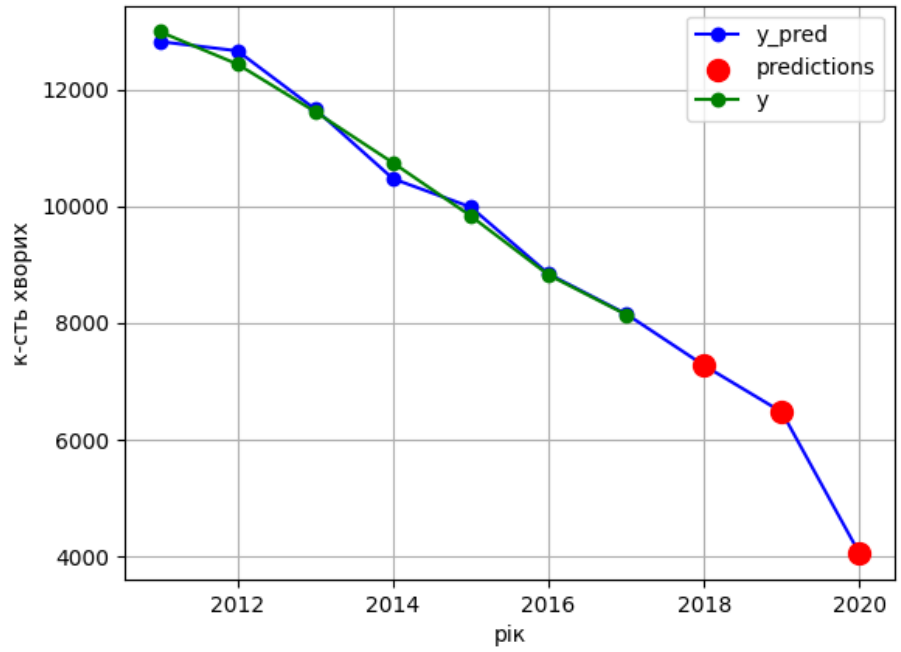



Рисунок 4.22 – Результат тестування нейронної мережі, що складається з чотирьох шарів (32, 16, 12, 8) та 500 епох

Дані для подальшого аналізу можна взяти з таблиці 3.2, але без стовпчика «ВВП». Розглянемо різні варіанти побудови нейромережі. Результати виходять наступними: двошарова з 16, 8 нейронами та 150 епох (рис. 4.23); двошарова з 16,8 нейронами та 500 епох (рис. 4.24); тришарова з 16, 12, 8 нейронами 150 епох (рис. 4.25); тришарова з 16, 12, 8 нейронами та 500 епох (рис. 4.26); чотирьохшарова з 32, 16, 12, 8 нейронами та 500 епох (рис. 4.27). За цими даними можна сказати, що на загальну тенденцію ВВП має невеликий вплив, але точність дещо впала в порівняні з першими варіантами, але при цьому можна сказати, що залежність прослідковується.

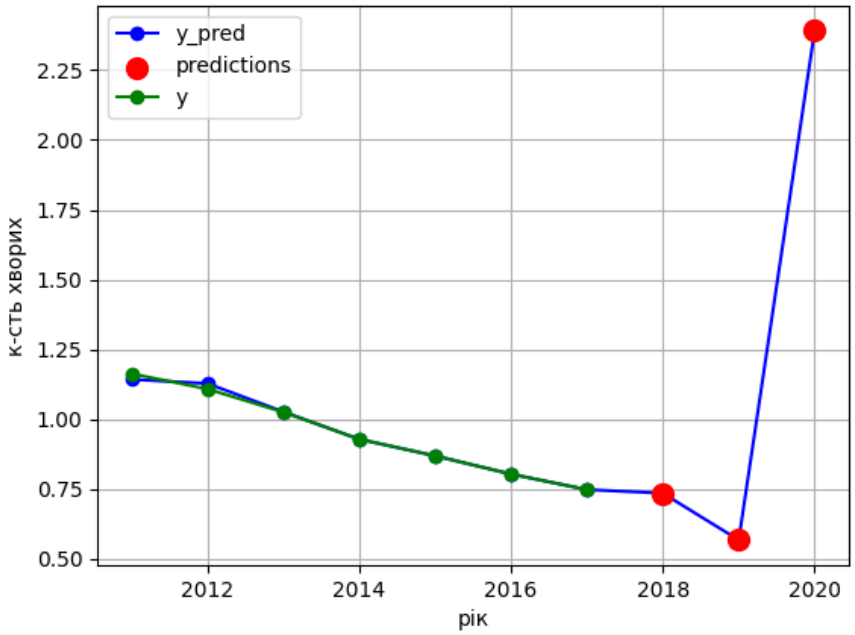



Рисунок 4.23 – Результат тестування нейронної мережі, що складається з двох шарів (16, 8) та 150 епох

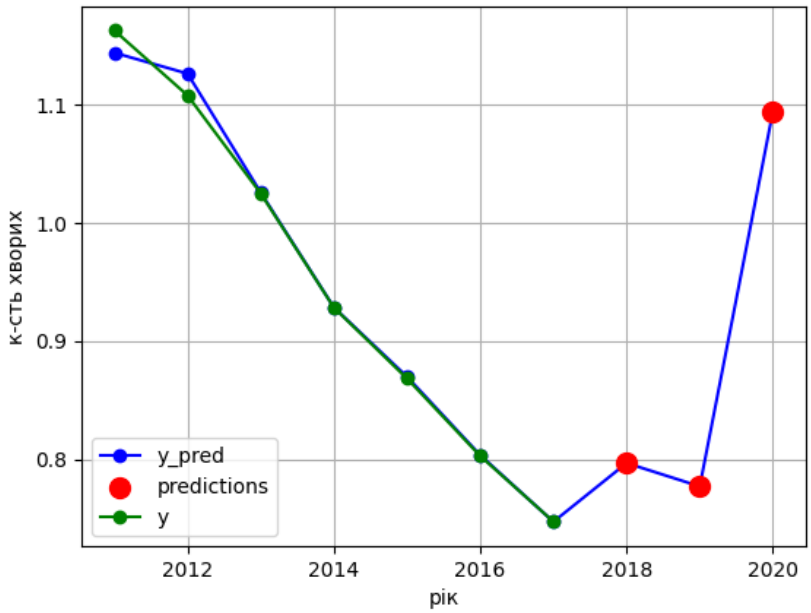



Рисунок 4.24 – Результат тестування нейронної мережі, що складається з двох шарів (16, 8) та 500 епох

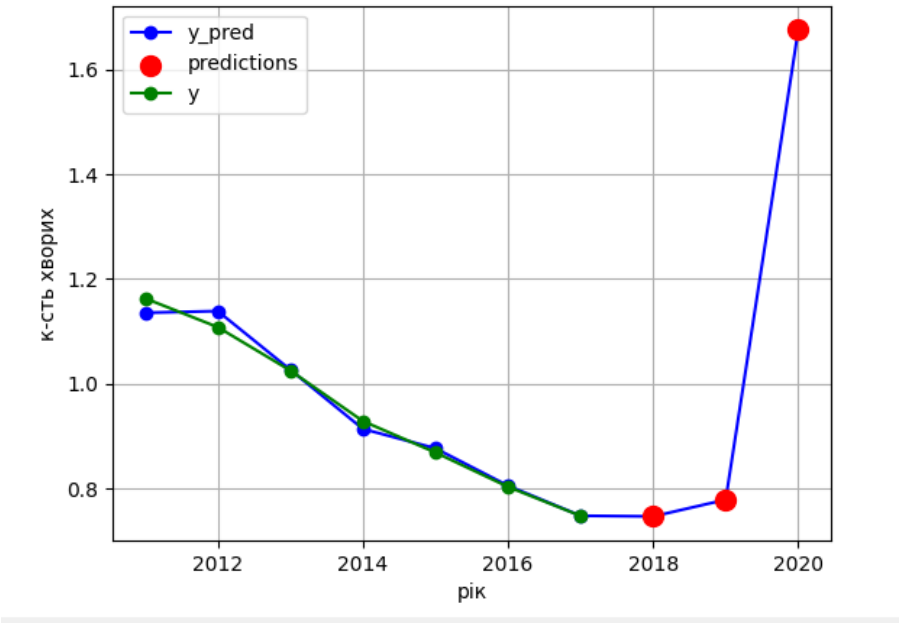



Рисунок 4.25 – Результат тестування нейронної мережі, що складається з трьох шарів(16, 12, 8) та 150 епох

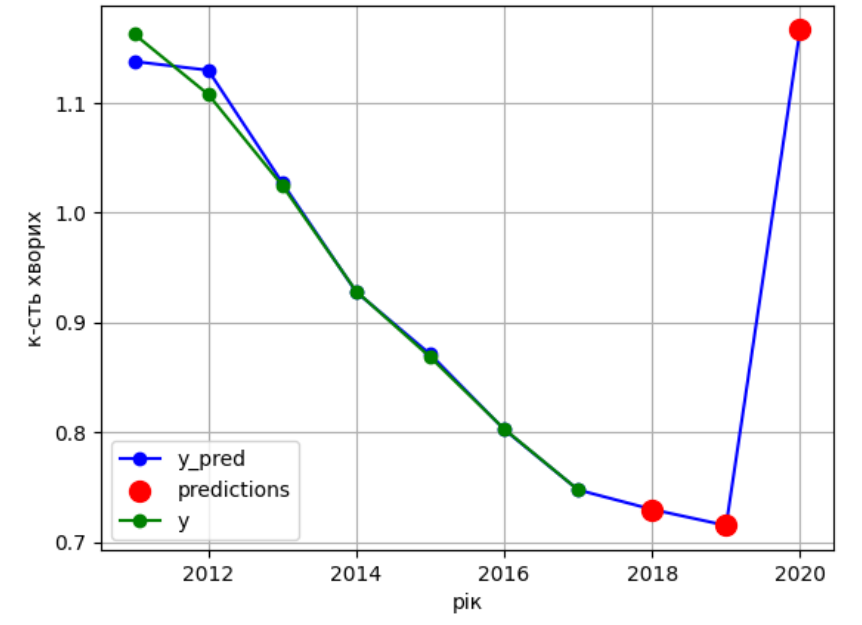



Рисунок 4.26 – Результат тестування нейронної мережі, що складається з трьох шарів (16, 12, 8) та 500 епох

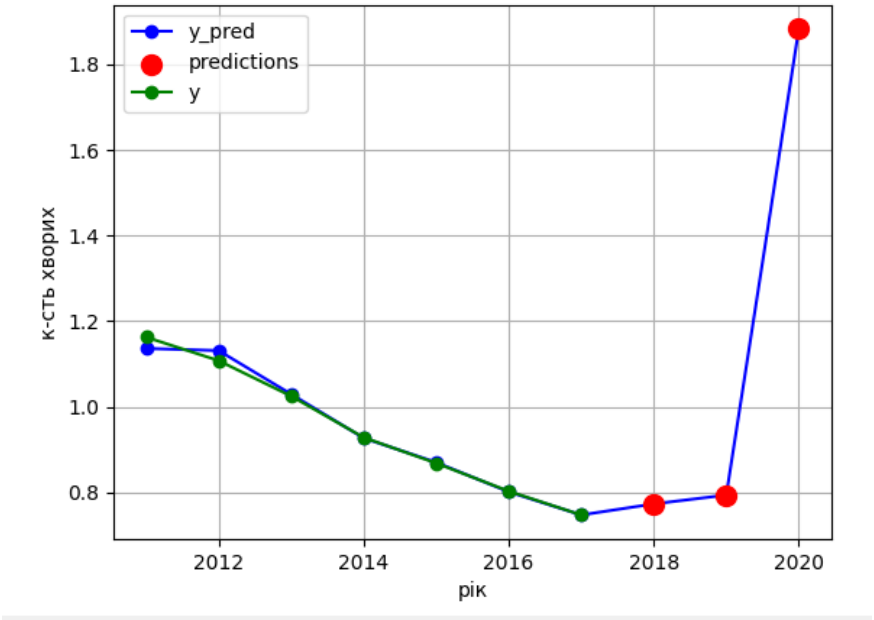



Рисунок 4.27 – Результат тестування нейронної мережі, що складається з чотирьох шарів (32, 16, 12, 8) та 500 епох

# ВИСНОВКИ

У даній роботі було досліджено вплив якості повітря та інших факторів на рівень захворюваності населення з використанням методів машинного навчання. Для побудови моделей прогнозування використовувалися дані про викиди забруднюючих речовин, використання відновлювальної енергії, внутрішній валовий продукт, бюджет Охорони Здоров'я та статистика захворюваності за декілька років.

Було реалізовано та протестовано кілька архітектур штучних нейронних мереж з різними параметрами: кількість шарів змінювалася від 2 до 4, кількість нейронів у шарах від 4 до 32, а кількість епох навчання від 150 до 1000. Найточніші результати, коли на графіках практично не було видно відмінностей між фактичними та прогнозованими даними, були досягнуті при конфігураціях: два шари з 16 і 12 нейронами та 500 епох, три шари з 16, 12 і 8 нейронами та 500 епох, три шари з 20, 16 і 12 нейронами та 500 епох, а також два шари з 16 і 8 нейронами та 150 епох у випадку використання трьох параметрів замість чотирьох. Це свідчить про те, що підвищення складності моделі не завжди гарантує кращу якість прогнозу, тому важливо знаходити баланс між кількістю шарів, нейронів та тривалістю навчання.

Також було встановлено, що попередня обробка даних зокрема нормалізація значень є дуже важливою для стабільного та ефективного навчання моделей. Нормалізовані цільові значення значно підвищують точність виконання програмного коду.

Отримані результати свідчать про наявність певної кореляції між забрудненням повітря та рівнем захворюваності, що при правильній обробці даних дозволяє використовувати подібні моделі для моніторингу й попередження зростання хвороб у майбутньому.

# ПЕРЕЛІК ПОСИЛАНЬ

1. Estill J. Health impacts and social costs associated with air pollution in larger urban areas of Ukraine [Електронний ресурс] / J. Estill. – 2022. – 32 p. – Режим доступу: https://www.undp.org/sites/g/files/zskgke326/files/2023-03/Health%20impacts%20and%20social%20costs%20associated%20with%20air%20pollution%20in%20larger%20urban%20areas%20of%20Ukraine%20%28ENG%29.pdf

2. Neff T. L. Public Health Impacts [Електронний ресурс] / T. L. Neff // The Social Costs of Solar Energy. – Pergamon, 1981. – С. 32–59. – Режим доступу: https://doi.org/10.1016/b978-0-08-026315-1.50010-9 (дата звернення: 31.05.2025).

3. The effect of national protest in Ecuador on PM pollution [Електронний ресурс] / R. Zalakeviciute [та ін.] // Scientific Reports. – 2021. – Т. 11, № 1. – Режим доступу: https://doi.org/10.1038/s41598-021-96868-6 (дата звернення: 31.05.2025).

4. Gradient Boosting Machine to Assess the Public Protest Impact on Urban Air Quality [Електронний ресурс] / R. Zalakeviciute [та ін.] // Applied Sciences. – 2021. – Т. 11, № 24. – С. 12083. – Режим доступу: https://doi.org/10.3390/app112412083 (дата звернення: 31.05.2025).

5. Gillespie R. D. Psychological Effects of War on Citizen and Soldier [Електронний ресурс] / R. D. Gillespie // The American Journal of the Medical Sciences. – 1942. – Т. 204, № 2. – С. 286. – Режим доступу: https://doi.org/10.1097/00000441-194208000-00023 (дата звернення: 31.05.2025).

6. War Impact on Air Quality in Ukraine [Електронний ресурс] / R. Zalakeviciute [та ін.] // Sustainability. – 2022. – Т. 14, № 21. – С. 13832. – Режим доступу: https://doi.org/10.3390/su142113832 (дата звернення: 31.05.2025).

7. Protopsaltis C. Air pollution caused by war activity [Електронний ресурс] / C. Protopsaltis // AIR POLLUTION 2012, A. Coruna, Spain, 16–18 трав. 2012 р. – Southampton, UK, 2012. – Режим доступу: https://doi.org/10.2495/air120091 (дата звернення: 31.05.2025).

8. Hays S. P. From Conservation to Environment: Environmental Politics in the United States Since World War Two [Електронний ресурс] / S. P. Hays // Environmental Review: ER. – 1982. – Т. 6, № 2. – С. 14. – Режим доступу: https://doi.org/10.2307/3984153 (дата звернення: 31.05.2025).

9. A systematic review and meta-analysis of human biomonitoring studies on exposure to environmental pollutants in Iran [Електронний ресурс] / M. Hadei [та ін.] // Ecotoxicology and Environmental Safety. – 2021. – Т. 212. – С. 111986. – Режим доступу: https://doi.org/10.1016/j.ecoenv.2021.111986 (дата звернення: 31.05.2025).

10. Rothschild R. Détente from the Air: Monitoring Air Pollution during the Cold War [Електронний ресурс] / R. Rothschild // Technology and Culture. – 2016. – Т. 57, № 4. – С. 831–865. – Режим доступу: https://doi.org/10.1353/tech.2016.0109 (дата звернення: 31.05.2025).

11. Premature Mortality in the Kingdom of Saudi Arabia Associated with Particulate Matter Air Pollution from the 1991 Gulf War [Електронний ресурс] / R. H. White [та ін.] // Human and Ecological Risk Assessment: An International Journal. – 2008. – Т. 14, № 4. – С. 645–664. – Режим доступу: https://doi.org/10.1080/10807030802235052 (дата звернення: 31.05.2025).

12. El-Shobokshy M. S. The impact of the gulf war on the Arabian environment–I. Particulate pollution and reduction of solar irradiance [Електронний ресурс] / M. S. El-Shobokshy, Y. G. Al-Saedi // Atmospheric Environment. Part A. General Topics. – 1993. – Т. 27, № 1. – С. 95–108. – Режим доступу: https://doi.org/10.1016/0960-1686(93)90074-9 (дата звернення: 31.05.2025).

13. Regional air pollution caused by a simultaneous destruction of major industrial sources in a war zone. The case of April Serbia in 1999 [Електронний ресурс] / Zorka B. Vukmirović [та ін.] // Atmospheric Environment. – 2001. – Т. 35, № 15. – С. 2773–2782. – Режим доступу: https://doi.org/10.1016/s1352-2310(00)00530-6 (дата звернення: 31.05.2025).

14. Sohrabi S. Burden of Disease Assessment of Ambient Air Pollution and Premature Mortality in Urban Areas: The Role of Socioeconomic Status and

Transportation [Електронний ресурс] / Soheil Sohrabi, Joe Zietsman, Haneen Khreis // International Journal of Environmental Research and Public Health. – 2020. – Т. 17, № 4. – С. 1166. – Режим доступу: https://doi.org/10.3390/ijerph17041166 (дата звернення: 31.05.2025).

15. WHO says 99% of world's population breathes poor-quality air [Електронний ресурс] // ABC News. – 2022. – Режим доступу: https://abcnews.go.com/Health/wireStory/99-worlds-population-breathes-poor-quality-air-83860782 (дата звернення: 30.05.2025).

16. Zhang W. The Clinical Significance of Recurrence of Papillary Thyroid Carcinoma [Електронний ресурс] / Wei Zhang, Xiaoyuan Zheng, Jian Zhao // Journal of Cancer. – 2015. – Т. 6, № 1. – С. 2–10. – Режим доступу: https://doi.org/10.3978/j.issn.2072-1439.2015.11.50.

17. EPA. Are you at risk from particles? How can particles affect your health? [Електронний ресурс] / Office of Air and Radiation. – EPA-452/F-03-001. – 2003. – С. 40–41. – Режим доступу: https://www.airnow.gov/sites/default/files/2018-03/pm-color.pdf (дата звернення: 31.05.2025).

18. Avenir Health. OneHealth Tool [Електронний ресурс]. – Режим доступу: https://www.avenirhealth.org/software-onehealth.php (дата звернення: 31.05.2025).

19. U.S. Agency for International Development (USAID). The DHS Program: Demographic and Health Surveys [Електронний ресурс]. – Режим доступу: https://dhsprogram.com/ (дата звернення: 31.05.2025).

20. University of Groningen, Groningen Growth and Development Centre. Penn World Table version 10.0 [Електронний ресурс]. – Режим доступу: https://www.rug.nl/ggdc/productivity/pwt/?lang=en (дата звернення: 31.05.2025).University of Groningen, Groningen Growth and Development Centre. Penn World Table version 10.0 [Internet]. Available from: https://www.rug.nl/ggdc/productivity/pwt/?lang=en.

21. Державна служба статистики України [Електронний ресурс]. – Режим доступу: https://ukrstat.gov.ua/ (дата звернення: 31.05.2025).

22. Офіційний вебпортал парламенту України. Законодавство України [Електронний ресурс]. – Режим доступу: https://zakon.rada.gov.ua/laws (дата звернення: 31.05.2025).

23. Фінансовий портал «Мінфін». Архів курсів валют [Електронний ресурс]. – Режим доступу: https://index.minfin.com.ua/exchange/archive (дата звернення: 31.05.2025).

24. Веб-портал «Доступ до правди». Запит «Захворюваність населення України» [Електронний ресурс]. – Режим доступу: https://dostup.org.ua/request/zakhvoriuvanist_nasieliennia_ukr (дата звернення: 31.05.2025).

25. Веб-портал для конвертації файлів PDF у Excel [Електронний ресурс]. – Режим доступу: https://pdfhouse.com/en/pdf-to-excel (дата звернення: 31.05.2025).

26. Населення України – Вікіпедія [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/Чисельність_населення_України (дата звернення: 31.05.2025).

27. Keras: Deep Learning for Humans – офіційна бібліотека [Електронний ресурс]. – Режим доступу: https://keras.io/ (дата звернення: 31.05.2025).

Додаток А

parsing.py

```
import requests
from bs4 import BeautifulSoup
from datetime import datetime, timedelta
from concurrent.futures import ThreadPoolExecutor, as_completed

def generate_dates(start_date, end_date):
    current = start_date
    while current <= end_date:
        yield current.strftime('%Y-%m-%d')
        current += timedelta(days=1)

def get_usd_rate(date_str):
    url = f" /{date_str}/"
    try:
        response = requests.get(url, timeout=10)
        response.raise_for_status()
    except Exception as e:
        return f"{date_str}: ошибка запроса ({e})"

    soup = BeautifulSoup(response.text, 'html.parser')
    table = soup.find('table', class_='zebra')
    if not table:
        return f"{date_str}: таблица не найдена"

    for row in table.find_all('tr'):
        cols = row.find_all('td')
```

```
        if len(cols) >= 6:
            code_num = cols[0].get_text(strip=True)
            code_txt = cols[1].get_text(strip=True)
            if code_num == '840' or code_txt == 'USD':
                currency = cols[3].get_text(strip=True)
                rate = cols[4].get_text(strip=True)
                change = cols[5].get_text(strip=True)
                change_pct = cols[6].get_text(strip=True) if len(cols) > 6 else ""
                return f"{date_str}: {code_txt} ({currency}) = {rate} грн | Δ {change} |
{change_pct}"
    return f"{date_str}: USD не найден"

start_date = datetime(2024, 1, 1)
end_date = datetime(2024, 12, 31)

dates = list(generate_dates(start_date, end_date))

results = []

with ThreadPoolExecutor(max_workers=16) as executor:
    future_to_date = {executor.submit(get_usd_rate, date): date for date in dates}
    for future in as_completed(future_to_date):
        result = future.result()
        print(result)
        results.append(result)

with open("usd_minfin_2024.txt", "w", encoding="utf-8") as f:
    for line in sorted(results):
        f.write(line + "\n")
```

Додаток Б

sort_data.py

```
import pandas as pd
import re

# === Завантаження всіх аркушів ===
file_path = 'data.xlsx'  # Заміни на свій шлях
xls = pd.ExcelFile(file_path)

all_data = []

# === Обробка кожного аркуша ===
for sheet_name in xls.sheet_names:
    df = xls.parse(sheet_name)

    # Пропускаємо пусті рядки
    df = df.dropna(subset=[df.columns[0]])

    # Виділення року та вікової групи з назви аркуша
    match = re.match(r"(\d{4})\s*\(([\d+–-]+)\)", sheet_name)
    if not match:
        continue
    year, age_group = match.groups()

    # Обробка колонок
    df_long = df.melt(id_vars=[df.columns[0]], var_name="column",
value_name="value")
    df_long.rename(columns={df.columns[0]: "Хвороба"}, inplace=True)
```

```
    # Парсинг категорії, статі, вікової підгрупи
    def parse_column(col_name):
        col_name = str(col_name).strip().lower()

        # Виправлення дати
        if re.match(r"\d{4}-\d{2}-\d{2}", col_name):
            if "07-14" in col_name or "7-14" in col_name:
                col_name += " 7-14"

        age_subgroup = ''
        gender = ''
        category = ''

        # Вікова підгрупа
        if '0-6' in col_name:
            age_subgroup = '0-6'
        elif '7-14' in col_name or '07-14' in col_name:
            age_subgroup = '7-14'
        elif '0-14' in col_name:
            age_subgroup = '0-14'
        elif '15-17' in col_name:
            age_subgroup = '15-17'
        elif '18' in col_name:
            age_subgroup = '18+'

        # Категорія — беремо повністю (навіть якщо з ; )
        known_keywords = ['виявлено під час профоглядів', 'вперше', 'диспансерн',
'усього', 'захворювань']
        if any(keyword in col_name for keyword in known_keywords):
```

```
            category = col_name  # лишаємо як є повністю
        else:
            category = 'невизначено'

        # Стать — тільки якщо нема згадки про вік у категорії
        if ';' in col_name and 'чоловіки' in col_name and 'жінки' in col_name and
'років' in col_name:
            gender = 'всі'
        elif 'чоловік' in col_name or 'хлопц' in col_name or 'юнаків' in col_name:
            gender = 'чоловіки'
        else:
            gender = 'всі'

        return age_subgroup, gender, category


    df_long[['Вікова підгрупа', 'Стать', 'Категорія']] =
df_long['column'].apply(parse_column).apply(pd.Series)

    df_long['Вікова група'] = age_group

    df_long['Рік'] = year

    df_long = df_long[['Хвороба', 'Рік', 'Вікова група', 'Вікова підгрупа',
'Категорія', 'Стать', 'value']]

    df_long = df_long.dropna(subset=['value'])

    all_data.append(df_long)
```

```
# === Об'єднання всього у tidy-формат ===
df_all = pd.concat(all_data, ignore_index=True)

# === Збереження у CSV ===
df_all.to_csv("усі_захворювання_tidied.csv", index=False, encoding='utf-8-sig')
```

Додаток В

search_anomal.py

```
import pandas as pd
import numpy as np

# Завантаження даних
df = pd.read_csv("усі_захворювання_tidied.csv")

# Перетворюємо на числові значення, на випадок якщо value був рядком
df["value"] = pd.to_numeric(df["value"], errors="coerce")

# Видаляємо рядки з пропущеними значеннями
df = df.dropna(subset=["value"])

# Створюємо ключ для групування
df["group_key"] = df[["Хвороба", "Вікова група", "Вікова підгрупа", "Категорія",
"Стать"]].astype(str).agg("|".join, axis=1)



# Список для збереження підозрілих значень
suspicious = []

# Перевірка кожної групи
for group, group_df in df.groupby("group_key"):
    if len(group_df) < 3:
        continue  # замало років для порівняння

    values = group_df["value"]
```

```
    for idx, row in group_df.iterrows():
        other_values = values.drop(idx)
        median_val = other_values.median()

        if median_val == 0:
            continue  # не можна ділити на 0

        ratio = row["value"] / median_val
        if ratio >= 9 or ratio <= 0.09:
            suspicious.append(row)

# Результат у DataFrame
suspicious_df = pd.DataFrame(suspicious)

# Збереження у файл
suspicious_df.to_csv("підозрілі_значення.csv", index=False, encoding='utf-8-sig')

print(f"Знайдено {len(suspicious_df)} підозрілих значень. Збережено у
'підозрілі_значення.csv'")
```

Додаток Г

normuvannya.py

```
import pandas as pd

# Завантаження даних
df = pd.read_csv("усі_захворювання_tidied.csv")

# Масив нормалізації
naselennya = [
    45778.5, 45633.6, 45553, 43087.7, 42928.9, 42760.5, 42584.5
]

# Перетворення 'Рік' у число (на випадок, якщо це рядок)
df['Рік'] = pd.to_numeric(df['Рік'], errors='coerce')

# Перетворення 'value' у число (рядки — у числа, пропуски — у NaN)
df['value'] = pd.to_numeric(df['value'], errors='coerce')

# Створення відповідності Рік → значення з naselennya
sorted_years = sorted(df['Рік'].dropna().unique())
year_to_naselennya = dict(zip(sorted_years, naselennya))

# Додаємо колонку з відповідним значенням нормалізації
df['naselennya_value'] = df['Рік'].map(year_to_naselennya)

# Обчислення нормованого значення (тільки якщо обидві величини числові)
df['Нормоване значення'] = df['value'] / df['naselennya_value']
```

```
# Зберігаємо результат
df.to_csv("усі_захворювання_tidied_нормовані.csv", index=False,  encoding='utf-8-sig')
```

Додаток Д

search_linear.py

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# === Завантаження даних ===
df = pd.read_csv("усі_захворювання_tidied_нормовані.csv", encoding="utf-8-sig")

# === Перетворення типів ===
df["Рік"] = pd.to_numeric(df["Рік"], errors="coerce")
df["Нормоване значення"] = pd.to_numeric(df["Нормоване значення"],
errors="coerce")

# === Унікальний ключ для кожної групи ===
cols = ["Хвороба", "Вікова група", "Вікова підгрупа", "Категорія", "Стать"]
df[cols] = df[cols].fillna("Невідомо").astype(str)
df["group_key"] = df[cols].agg("|".join, axis=1)

# === Відбір тільки повних груп (по 7 років) ===
group_counts = df.groupby("group_key")["Рік"].nunique()
valid_keys = group_counts[group_counts == 7].index
df_filtered = df[df["group_key"].isin(valid_keys)]

# === Обчислення тренду, R² і аномалії ===
```

```
trend_data = []

for key, group in df_filtered.groupby("group_key"):
    group = group.sort_values("Рік")
    subgroup = group.dropna(subset=["Рік", "Нормоване значення"])

    if len(subgroup) < 2:
        continue  # недостатньо даних для моделі

    X = subgroup[["Рік"]]
    y = subgroup["Нормоване значення"]

    model = LinearRegression()
    model.fit(X, y)

    y_pred = model.predict(X)
    r2 = r2_score(y, y_pred)             # якість прилягання
    anomaly = np.sqrt(np.mean((y - y_pred) ** 2))  # RMSE

    trend_data.append({
        "group_key": key,
        "r2": r2,
        "anomaly_score": anomaly,
        "coef": model.coef_[0],
        "intercept": model.intercept_
    })

trend_df = pd.DataFrame(trend_data)

# === Збереження результатів ===
```

```
trend_df.sort_values("anomaly_score",
ascending=False).head(4).to_csv("топ_4_аномалії.csv", index=False,
encoding="utf-8-sig")
trend_df.sort_values("r2", ascending=False).head(4).to_csv("топ_4_лінійність.csv",
index=False, encoding="utf-8-sig")

# === Функція побудови графіка ===
def plot_group(df, group_key, ax):
    group = df[df["group_key"] == group_key].sort_values("Рік")
    sns.lineplot(data=group, x="Рік", y="Нормоване значення", marker="o", ax=ax)
    ax.set_title(group_key, fontsize=9)
    ax.tick_params(axis='x', rotation=45)

# === Побудова графіків ===

# Топ-4 за аномальністю
top_anomaly = trend_df.sort_values("anomaly_score", ascending=False).head(4)
fig, axes = plt.subplots(2, 2, figsize=(12, 8))
fig.suptitle("Топ-4 за аномальністю", fontsize=14)
for ax, key in zip(axes.flatten(), top_anomaly["group_key"]):
    plot_group(df_filtered, key, ax)
plt.tight_layout()
plt.show()

# Топ-4 за лінійністю (R²)
top_r2 = trend_df.sort_values("r2", ascending=False).head(4)
fig, axes = plt.subplots(2, 2, figsize=(12, 8))
fig.suptitle("Топ-4 за лінійністю (R²)", fontsize=14)
for ax, key in zip(axes.flatten(), top_r2["group_key"]):
    plot_group(df_filtered, key, ax)
```

```
plt.tight_layout()
plt.show()
```

Додаток Е

main.py

```
import pandas as pd
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import Input
from tensorflow.keras.optimizers import Adam
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
import numpy as np


dataset = pd.read_csv('data.txt', delimiter='\t', header=None)

X = dataset.iloc[:, :4]
y = dataset.iloc[:, -1]
y = y.head(7)
x1 = X.head(10)

y = y.to_numpy().reshape(-1, 1)

scaler = StandardScaler()
x1 = scaler.fit_transform(x1)

new_data = x1[7:10]
x1 = x1[:7]
```

```
model = Sequential()
model.add(Input(shape=(4,)))
model.add(Dense(16, activation='relu'))
model.add(Dense(12, activation='relu'))
model.add(Dense(1))

learning_rate = 0.05
optimizer = Adam(learning_rate=learning_rate)
model.compile(loss='mean_squared_error', optimizer=optimizer, metrics=['mae'])

model.fit(x1, y, epochs=150, batch_size=16)

y_pred = model.predict(x1)
mse = mean_squared_error(y, y_pred)
print(f"Mean Squared Error on Test Data: {mse:.2f}")

predictions = model.predict(new_data)

print("Predictions for new data:", predictions)



y = np.array(y.flatten())
combined_array = np.concatenate((y_pred.flatten(), predictions.flatten()))
indices = np.arange(2011, 2021)
plt.plot(indices, combined_array, marker='o', linestyle='-', color='b', label="y_pred")
plt.scatter(indices[-3:], combined_array[-3:], color='r', s=100, zorder=5,
label="predictions")
plt.plot(indices[:7], y, marker='o', linestyle='-', color='g', label="y")
plt.xlabel("рік")
plt.ylabel("к-сть хворих")
```

```
plt.title("")
plt.legend()
plt.grid(True)
plt.show()
```