ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ С.КУЗНЕЦЯ

КАФЕДРА ІНФОРМАТИКИ ТА КОМП'ЮТЕРНОЇ ТЕХНІКИ

Дипломний проєкт на тему: «Розроблення інформаційного модуля оцінки епідемічної небезпеки поширення COVID-19/ Development of an information module for the epidemic risk for the spread of COVID-19»

Виконав: студент 4 курсу, групи 6.04.126.010.19.1

ОП «Інформаційні системи та технології» Артем ДОЛГИЙ

Керівник: к.т.н., доцент Ольга ТЮТЮНИК

Об'єктом дослідження є інформаційні системи та технології оброблення даних для оцінки епідеміологічної небезпеки поширення COVID-19.

Предметом дослідження є інформаційний модуль оцінки епідемічної небезпеки поширення COVID-19.

Метою дипломного проекту є розроблення інформаційного модулю оцінки епідемічної небезпеки поширення COVID-19 та аналіз найбільш оптимальних моделей для прогнозування кількості пацієнтів в лікарнях.

Методи розроблення. Методи системного аналізу, статистичні методи обробки інформації, методи прогнозування, методи машинного навчання.

Актуальність предметної області

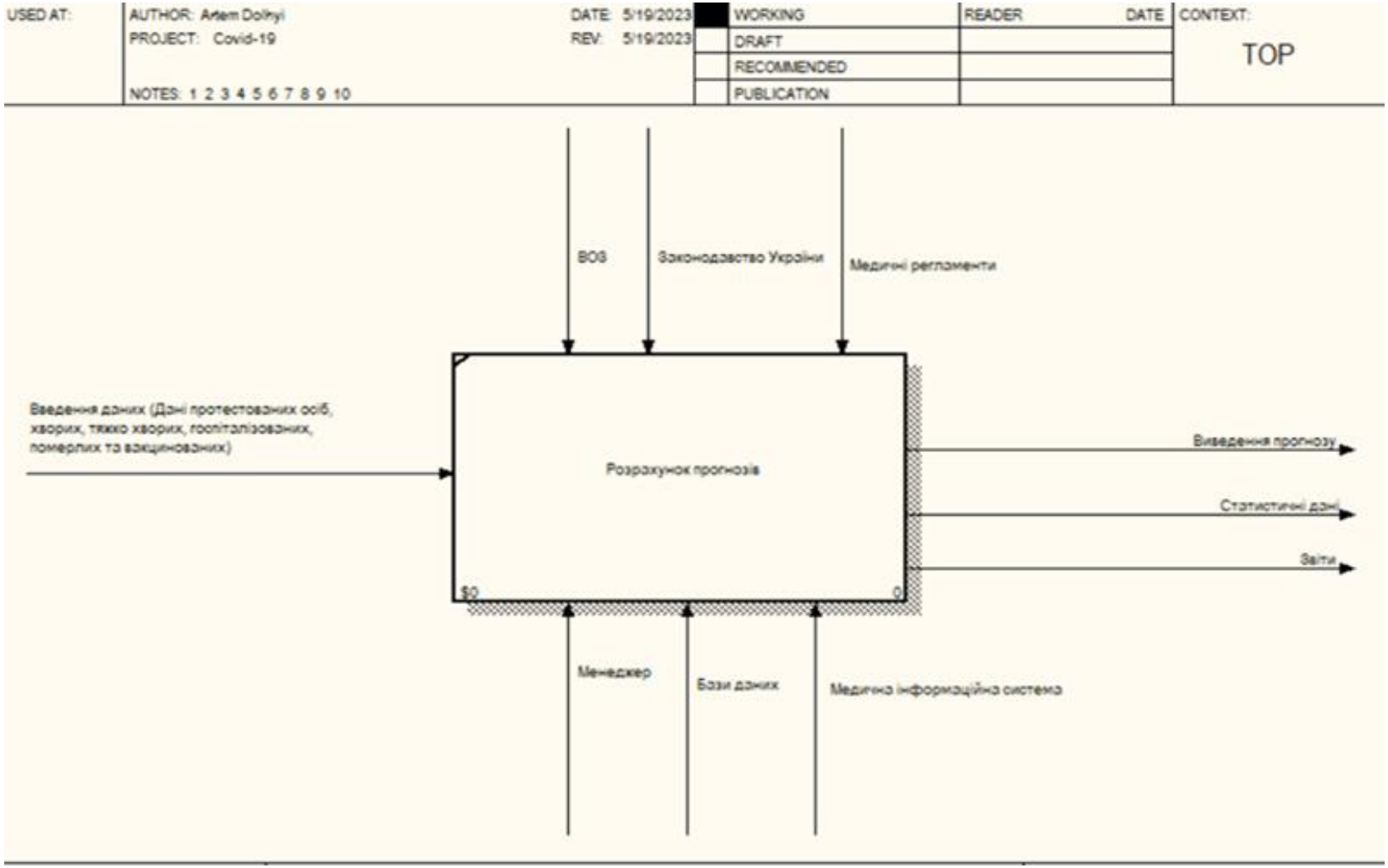
- Коронавіруси – це група вірусів, які здебільшого спричиняють незначні проблеми, що супроводжуються такими симптомами, як кашель та застуда. Більшість коронавірусів є нешкідливими для людей. Проте новий коронавірус COVID-19 є доволі агресивним.
- COVID-19 поширюється швидше, ніж інші вірусні ГРВІ.
- COVID-19 визнають дуже загрозливим тому, що при ураженні легень COVID-19 людина може не мати симптомів захворювання довгий час, але зміни у дихальних мережах хворого у цей час можуть бути вже не виправними.
- До недавнього часу не існувало вакцин та противірусних препаратів для лікування хвороби.
- Хоча показник летальності від COVID-19 становить всього кілька відсотків, пов'язана з цим пандемія викликала набагато більше смертей у всьому світі. За даними інформаційної моделі ВООЗ за три роки зафіксовано підтверджених випадків COVID-2019 більше 766 мільйонів та майже 7 мільйонів смертей, але фактична кількість може бути в кілька разів вище.
- За період пандемії імунітет населення зростає, завдяки вакцинації, знижуються літальні випадки, вже понад рік пандемія має низхідну тенденцію, що послаблює тиск на системи охорони здоров'я країн. У результаті чого, більшість країн повернулася до звичайного до пандемічного життя.
- 4 травня 2023 р. Всесвітня організація з охорони здоров'я оголосила про закінчення пандемії COVID-19.

Моделі машинного навчання

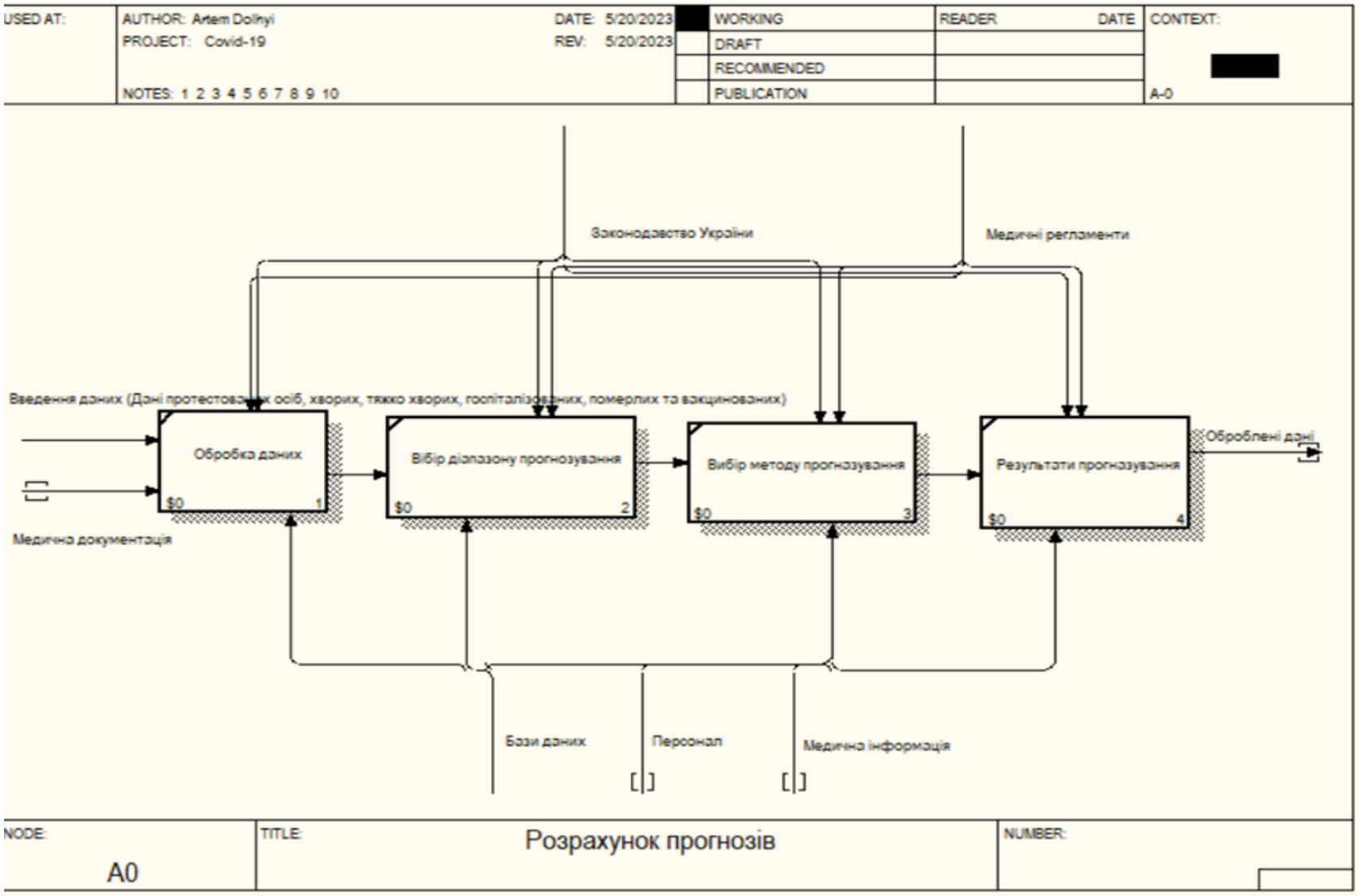
Для аналізу та прогнозування розвитку пандемії застосовуються наступні моделі:

- ARIMA (Авторегресивне інтегроване ковзне середнє);
- LTSM (довгострокова короткочасна пам’ять);
- Decision Forest (Ліс рішень)

Функціональна модель системи «Розрахунок прогнозів»



Декомпозиція роботи 1-го рівня «Обробка даних»



Постановка задач дослідження

Для розроблення інформаційного модулю оцінки епідемічної небезпеки поширення COVID-19 необхідно виконати наступні завдання, а саме:

- Підготувати та проаналізувати доступні набори даних;
- Розробити ARIMA, LSTM та Decision Forest для прогнозування кількості пацієнтів в лікарнях;
- Навчити моделі на наявних наборах даних;
- Порівняти та проаналізувати отримані результати прогнозування моделей;
- Визначити переваги та недоліки кожної з моделей прогнозування для кількості пацієнтів в лікарнях;
- Обрати найбільш оптимальну модель для використання в інформаційному модулі оцінки епідемічної небезпеки поширення COVID-19.

Аналіз ефективності застосування моделі ARIMA для оцінки епідемічної небезпеки поширення COVID-19

ARIMA ("Авторегресійна інтегрована ковзна середня") - це потужна статистична модель, яка широко використовується для прогнозування часових рядів.

Основна ідея ARIMA полягає в розкладанні часового ряду на три компоненти:

- авторегресію (AR);
- інтеграцію (I);
- ковзну середню (MA).

Комбінуючи ці компоненти, модель ARIMA може точно прогнозувати майбутні значення часового ряду.

Аналіз ефективності застосування моделі LSTM для оцінки епідемічної небезпеки поширення COVID-19

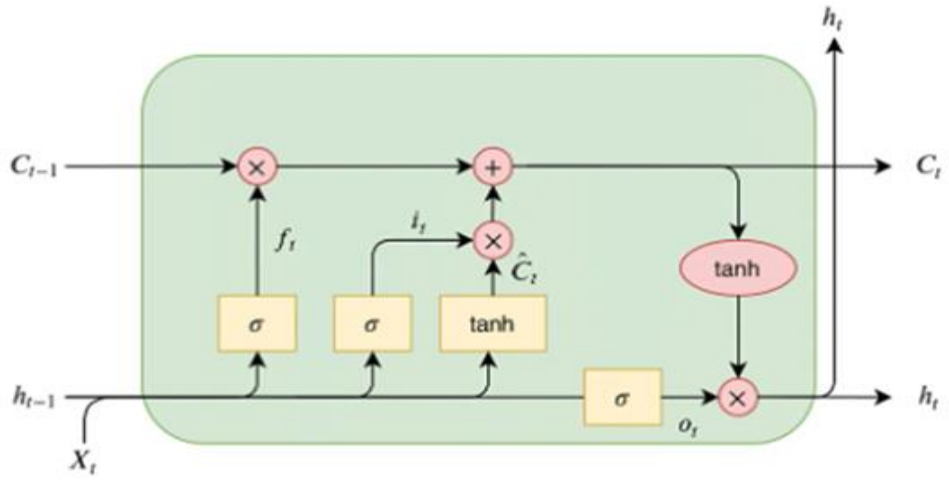
LSTM (LongShort-TermMemory) - це тип рекурентної нейронної мережі (RNN), яка була спеціально розроблена для вирішення проблеми зникнення та вибуху градієнту, що виникають при навчанні RNN на довгих послідовностях даних.

Модель LSTM включає в себе комірку пам'яті і троє воріт:

- ворота забуття;
- вхідні ворота;
- вихідні ворота.

Ці компоненти керують внутрішнім потоком інформації, дозволяючи моделі вибірково зберігати або відкидати не потрібну інформацію на різних часових кроках.

Ворота - це обчислювальний механізм, який регулює прохід інформації, фактично виступаючи як перемикач, який визначає кількість інформації, яка передається на кожному часовому кроці.



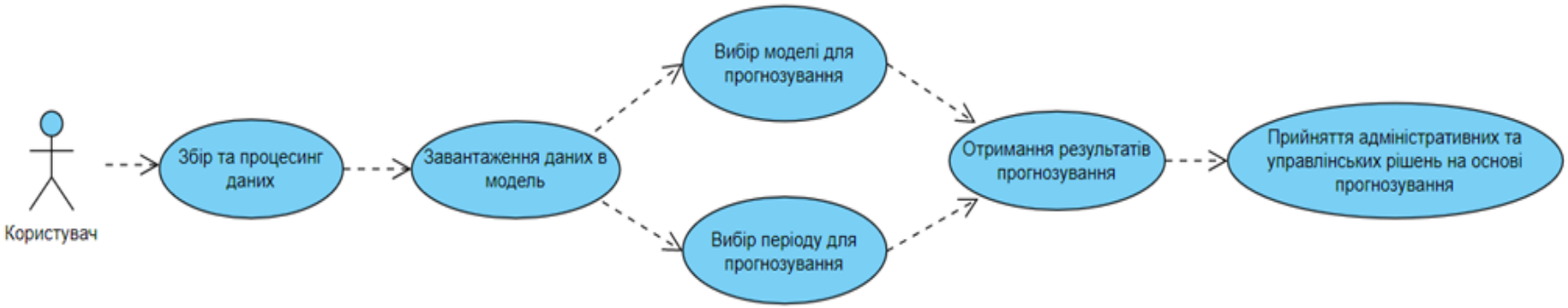
Аналіз ефективності застосування моделі Decision Forest для оцінки епідемічної небезпеки поширення COVID-19

Decision Forest (Ліс рішень) - є непараметричним контрольованим методом навчання, який використовується для класифікації та регресії.

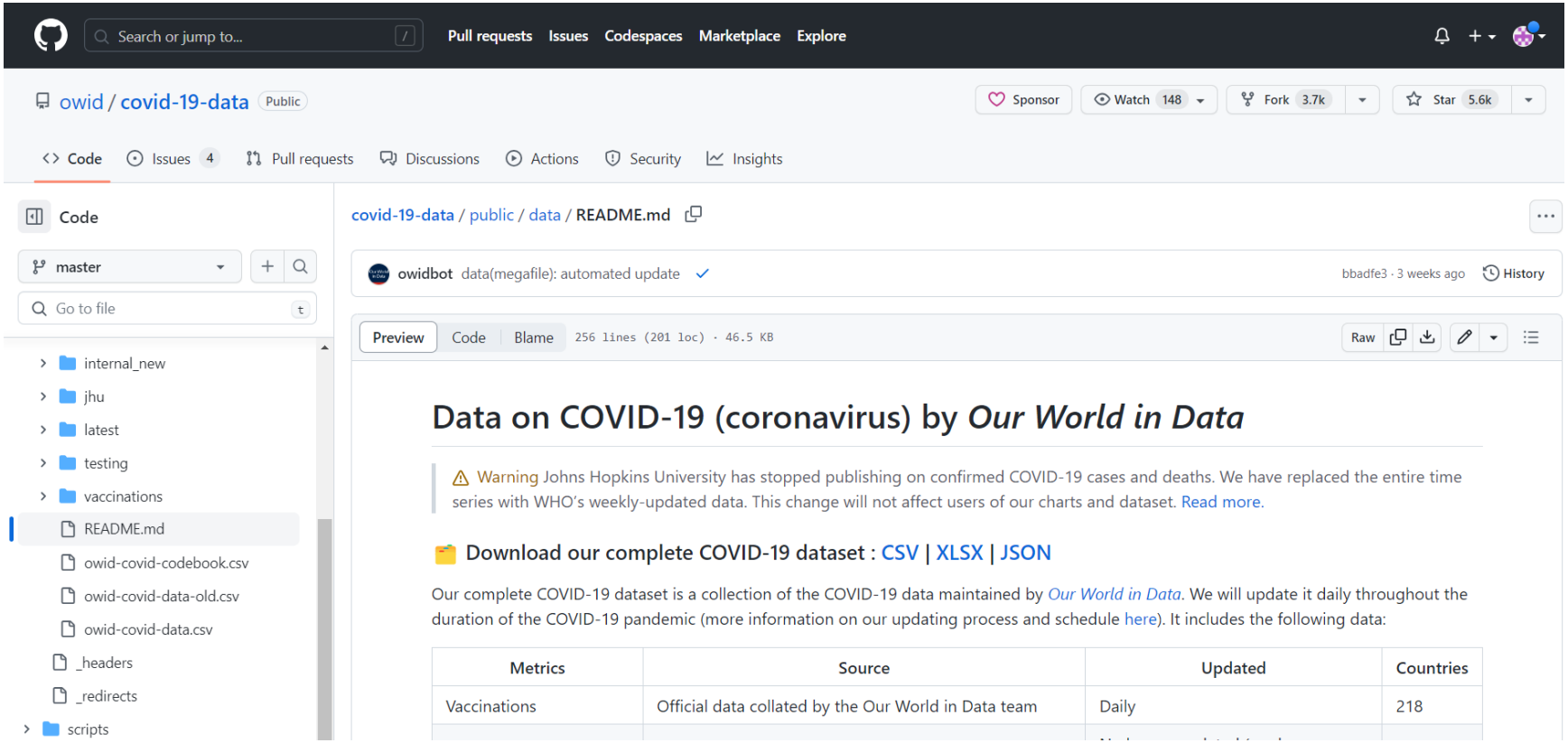
Decision Forest є ансамблем моделей рішень, який використовується для аналізу та прогнозування даних в різних областях, включаючи прогнозування часових рядів.

Decision Forest базується на понятті дерева рішень, де кожен вузол представляє розгалуження на основі певного атрибуту. У Decision Forest створюється багато дерев рішень, а потім прогнози об'єднуються шляхом голосування або усереднення для отримання остаточного прогнозу.

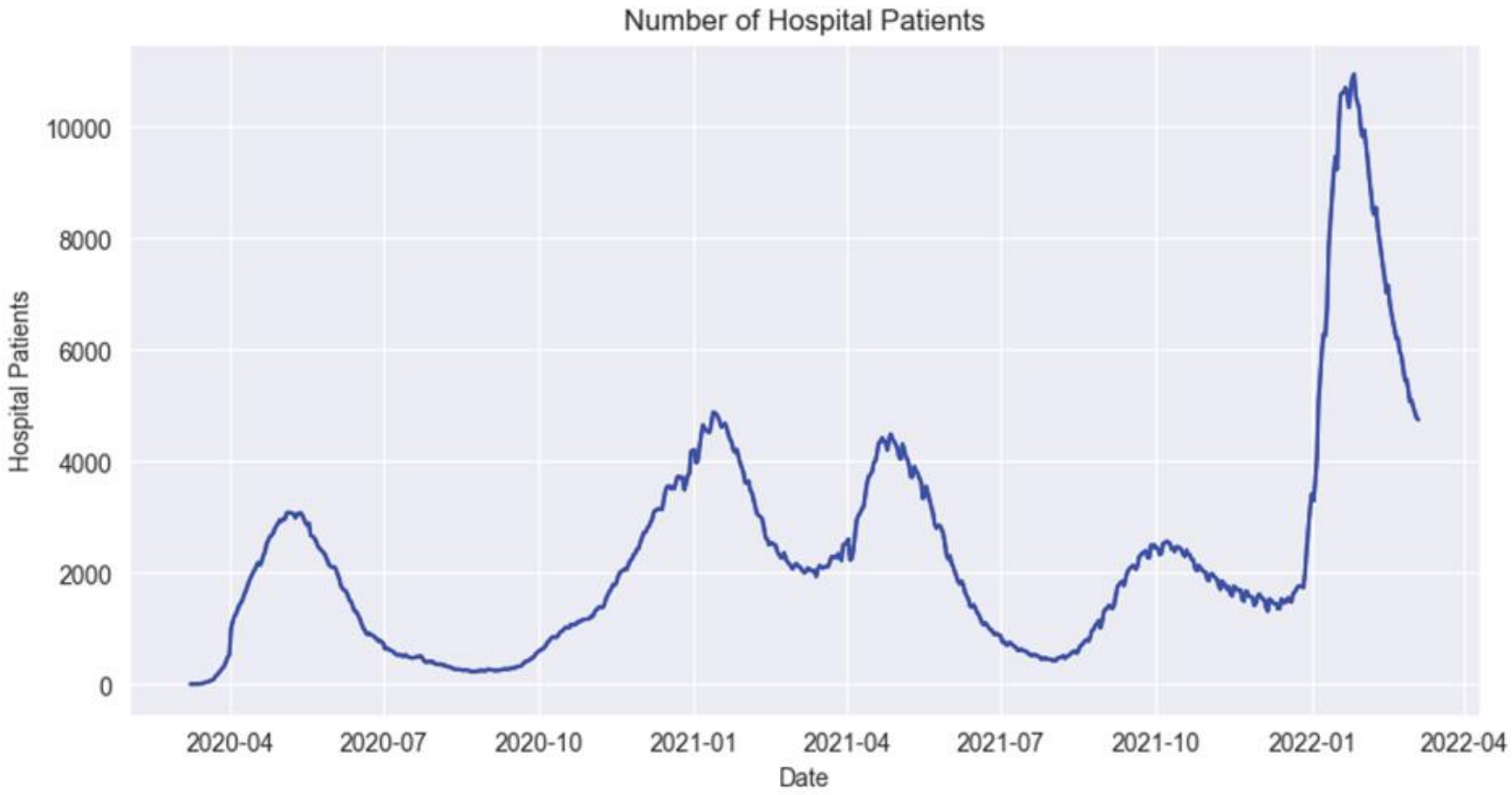
Діаграма варіантів використання інформаційного модуля оцінки епідемічної небезпеки поширення covid-19



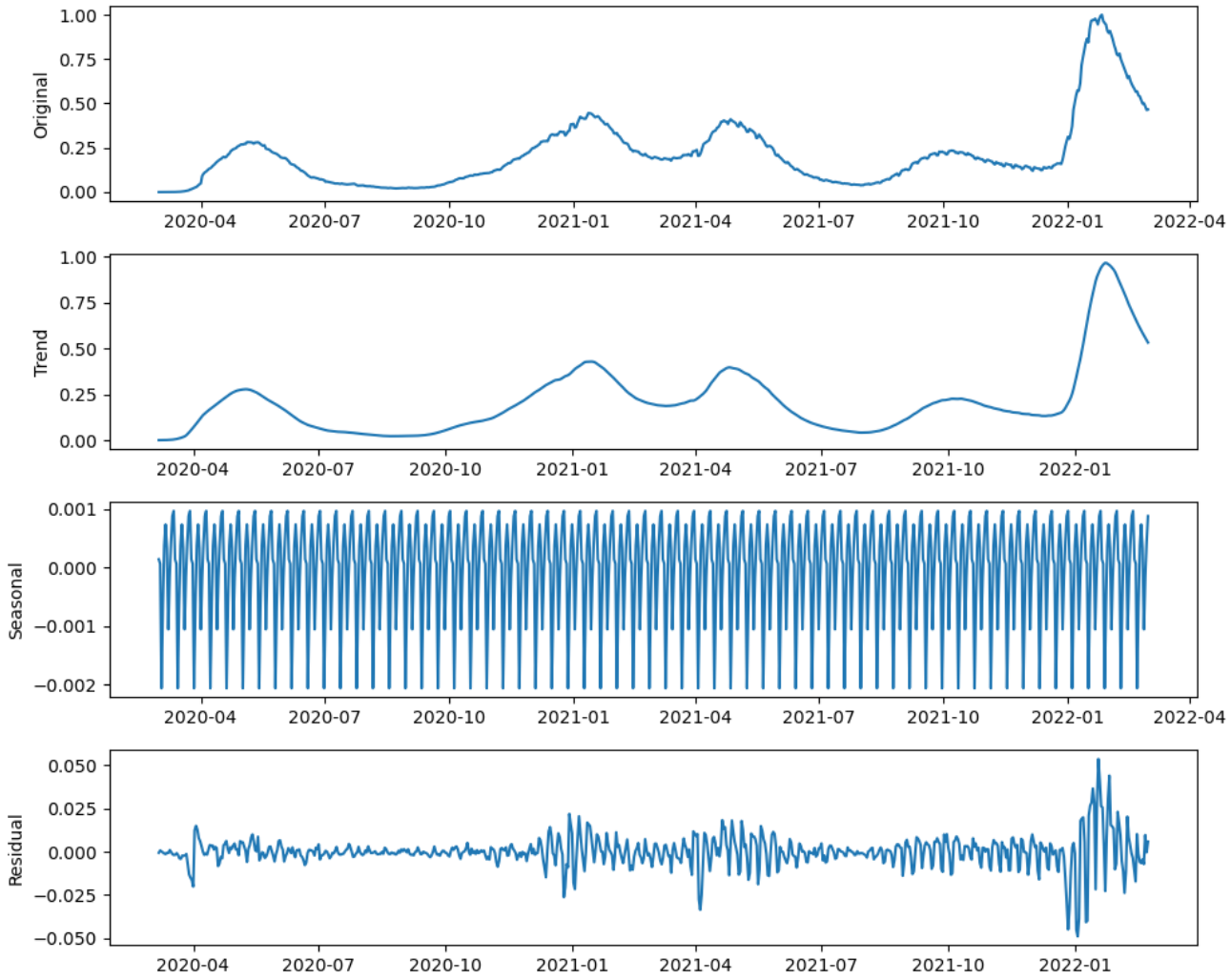
Джерело даних отримано з відкритого репозиторія “Data on COVID-19 by Our World in Data”



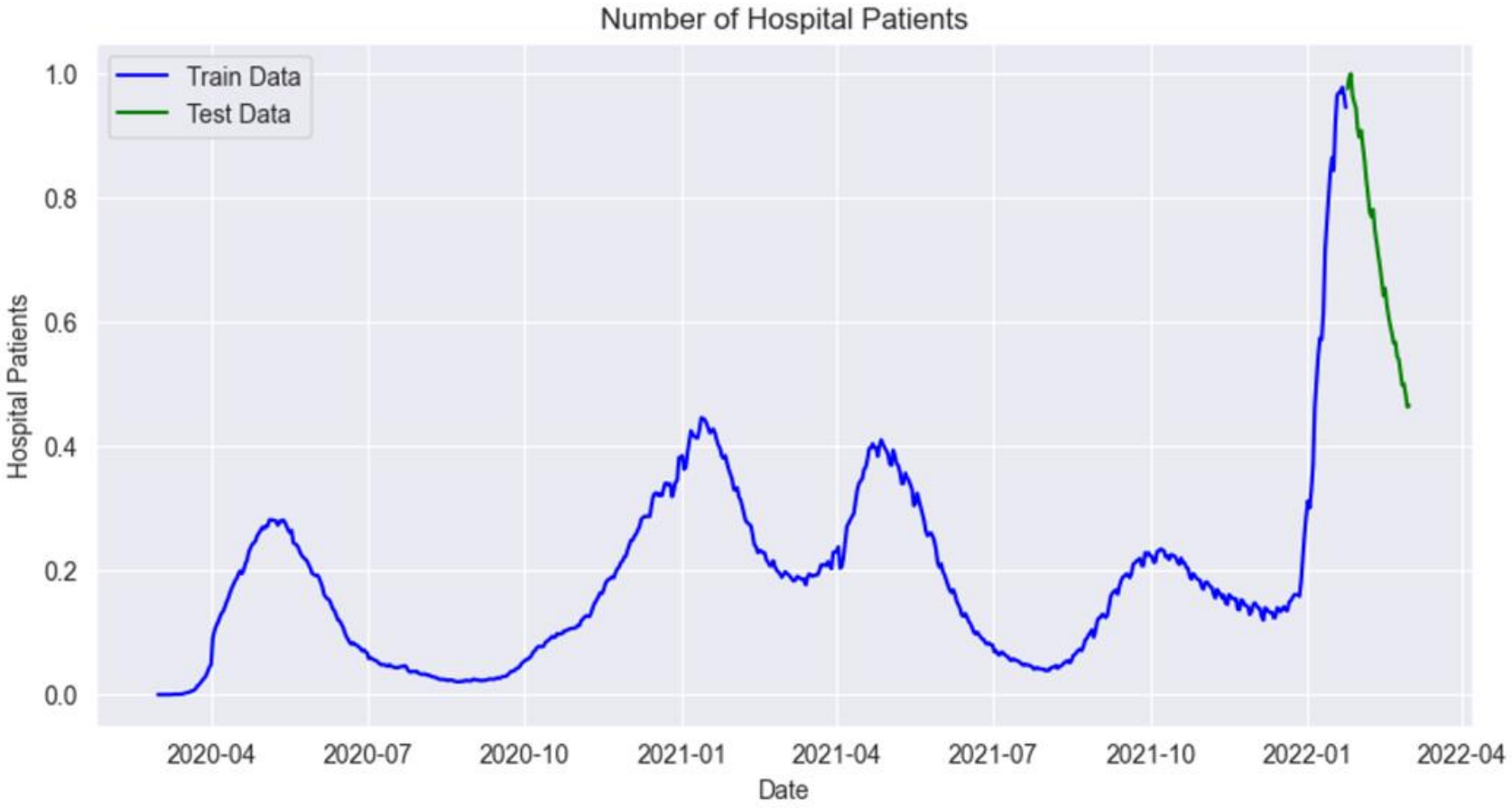
Дані про кількість госпіталізованих пацієнтів



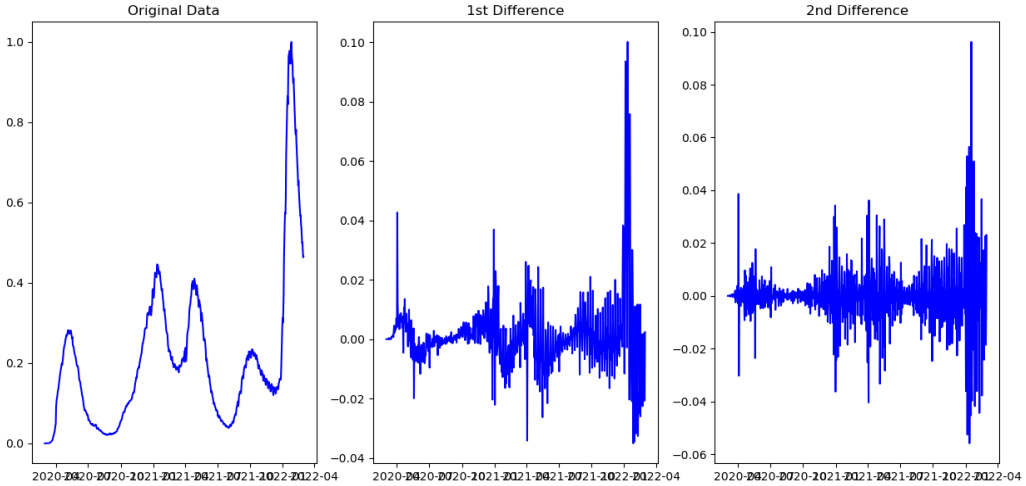
Сезонна декомпозиція набору даних

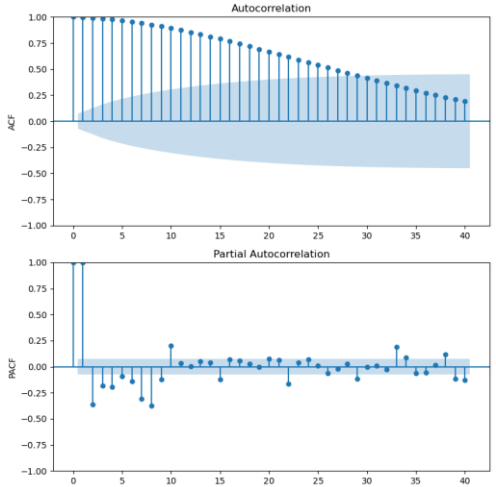


Розбиття даних на тренувальні та тестувальні набори

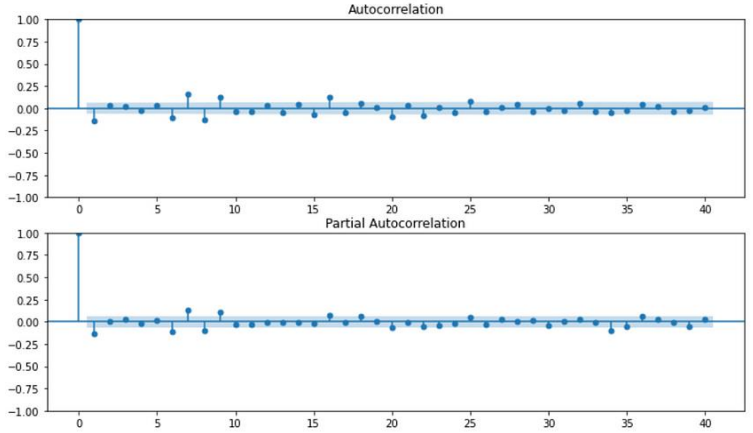


Результати пошуку оптимальних параметрів для ARIMA моделі

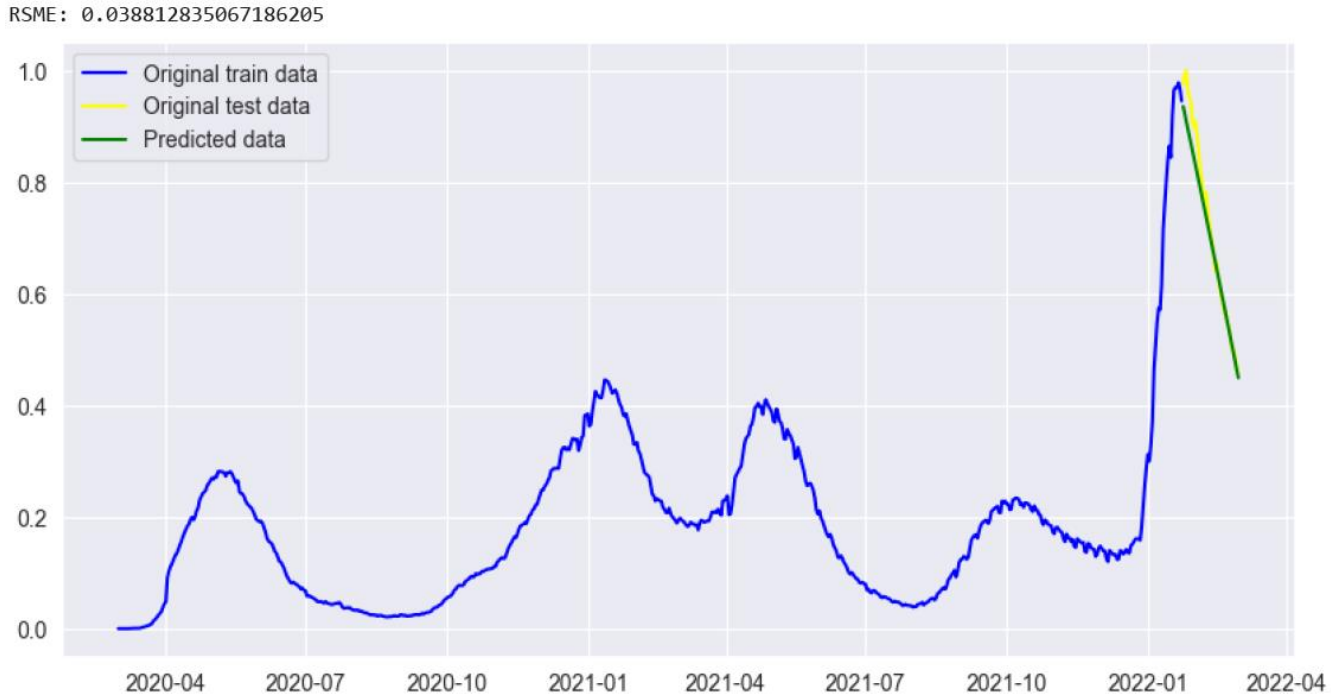





```
==============================================================================
Dep. Variable:          hosp_patients   No. Observations:                  694
Model:                 ARIMA(2, 3, 1)   Log Likelihood                2219.947
Date:                Tue, 23 May 2023   AIC                          -4431.894
Time:                        04:09:28   BIC                          -4413.741
Sample:                    03-01-2020   HQIC                         -4424.872
                         - 01-23-2022
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4333      0.016    -26.761      0.000      -0.465      -0.402
ar.L2         -0.3431      0.017    -20.506      0.000      -0.376      -0.310
ma.L1         -0.9993      0.044    -22.805      0.000      -1.085      -0.913
sigma2      9.358e-05   3.44e-06     27.201      0.000    8.68e-05       0.000
===================================================================================
Ljung-Box (L1) (Q):                   2.93   Jarque-Bera (JB):              5719.06
Prob(Q):                              0.09   Prob(JB):                         0.00
Heteroskedasticity (H):               9.39   Skew:                             1.32
Prob(H) (two-sided):                  0.00   Kurtosis:                        16.85
===================================================================================
```



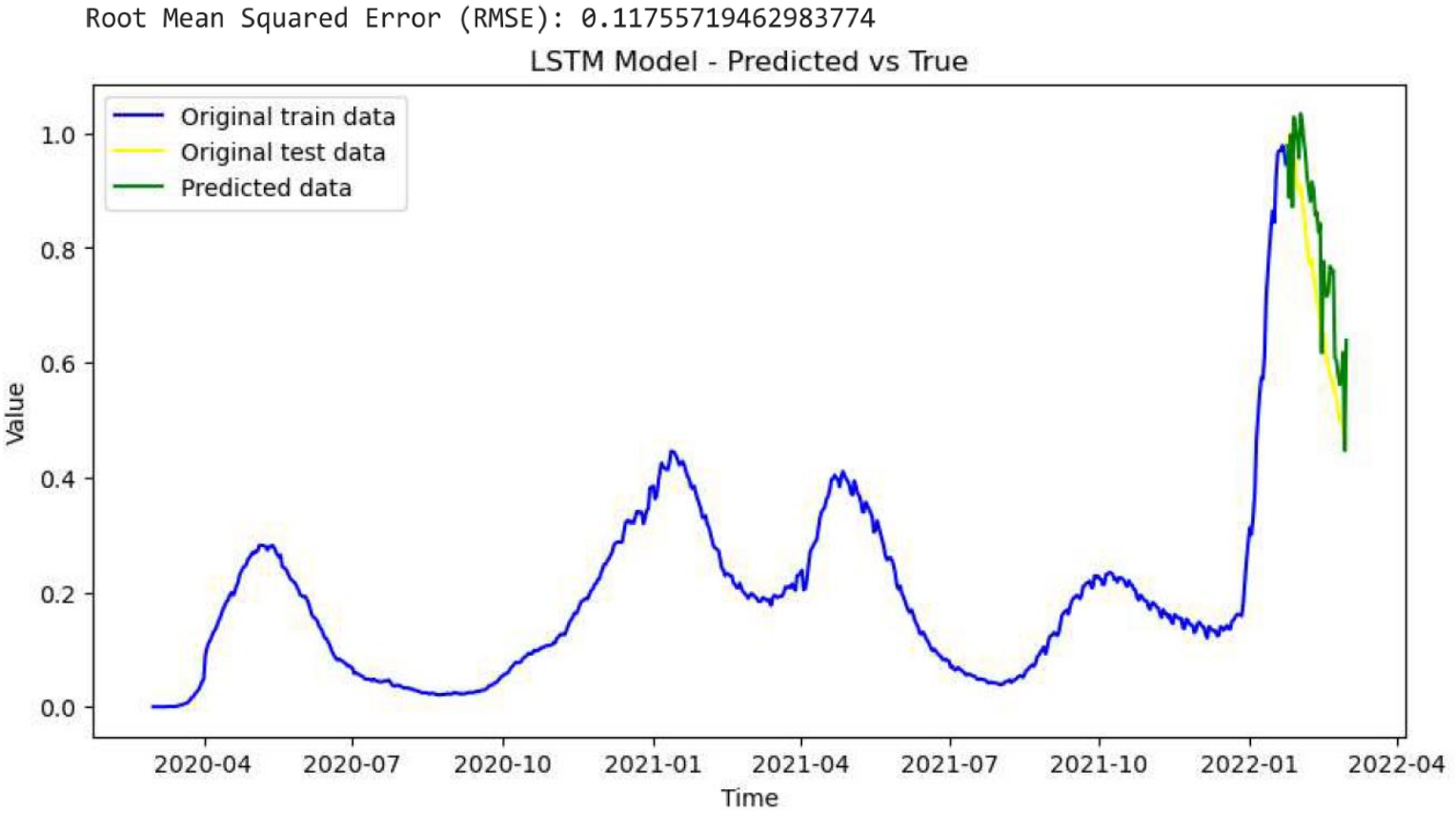
Результати прогнозування ARIMA моделі



RMSE (RootMeanSquaredError) - це метрика, що використовується для оцінки точності моделі прогнозування. Вона вимірює середньоквадратичну помилку між прогнозованими значеннями і фактичними значеннями в тестовому наборі даних.

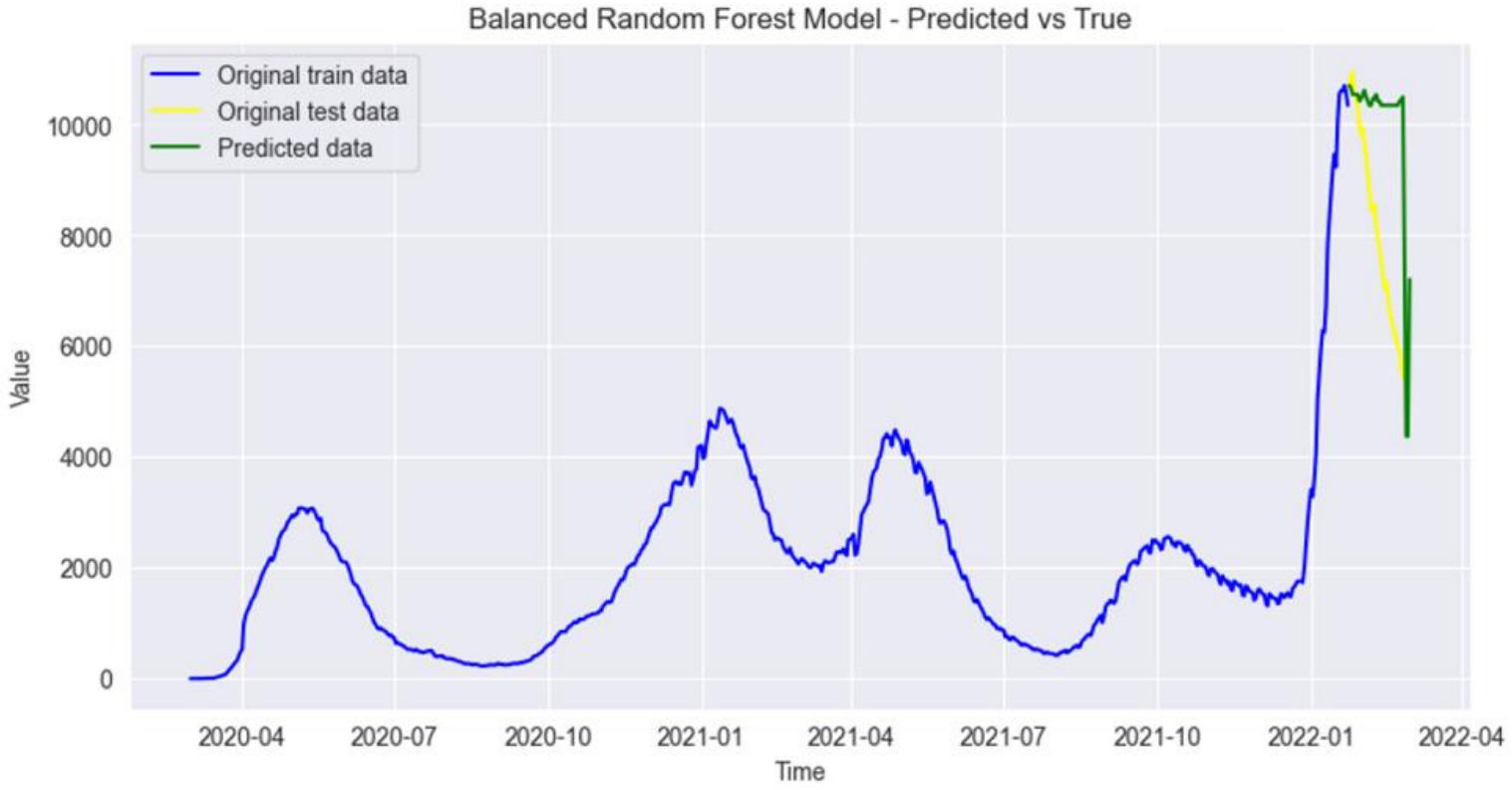
Значення RMSE дорівнює 0.0388, що свідчить про досить низьку середньоквадратичну помилку прогнозу моделі ARIMA. Це означає, що прогнозовані значення досить точно відповідають тестовим даним і модель має високу точність у прогнозуванні кількості пацієнтів в лікарнях.

Результати прогнозування LTSM моделі



Після виконання прогнозу з використанням LTSM моделі, отримано графік зі значенням RMSE: 0.1175. В даному випадку, отримане значення RMSE свідчить про помірну середньоквадратичну помилку прогнозу моделі LTSM.

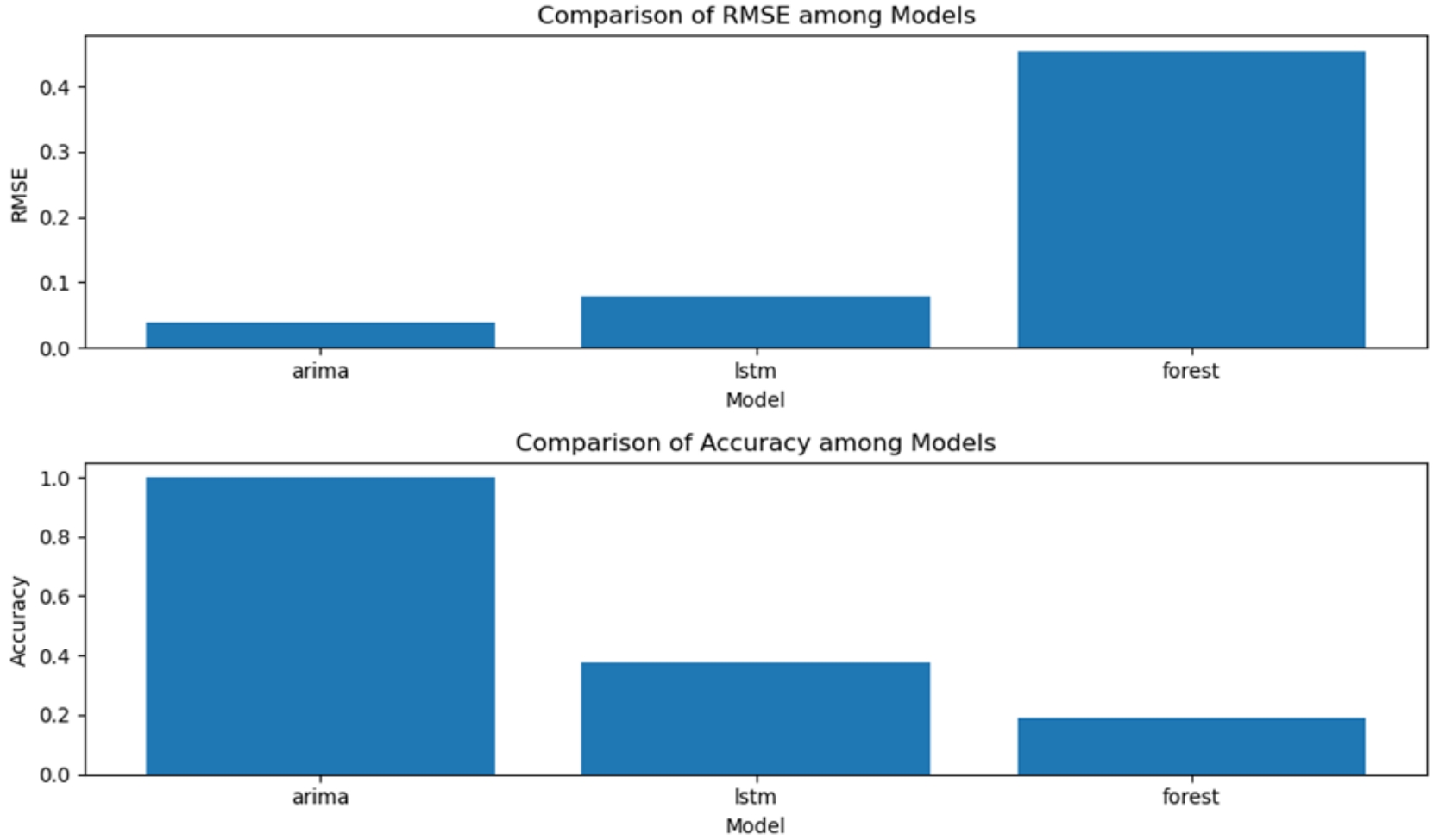
Результати прогнозування Decision Forest моделі



Root Mean Squared Error (RMSE): 0.4547771994013899

Отриманий графік для моделі Balanced Random Forest Classifier показав, що RootMeanSquaredError (RMSE) дорівнює 0.4547. Це свідчить про те, що модель не досягла оптимальної точності у прогнозуванні кількості пацієнтів в лікарнях.

Порівняльний аналіз результатів моделей



Результати дослідження

Тютюник В. В. Особливості прийняття експертами ситуаційного центру управлінських антикризових рішень в умовах епідемічної небезпеки поширення COVID-19 / В. В. Тютюник, О. О. Тютюник, А. О. Долгий // Запобігання виникненню надзвичайних ситуацій, реагування та ліквідація їх наслідків : матеріали круглого столу (вебінару), 23 лютого 2023 р. – Харків, 2023.– С. 150-152.

Висновки

Під час проведеного дослідження було проаналізовано три моделі прогнозування даних для COVID-19 та отримано наступні результати:

1. ARIMA модель - показала один з найкращих результатів прогнозування. Модель ARIMA вдалося точно відтворити тенденції та рухи у графіку кількості пацієнтів в лікарнях, що свідчить про її високу точність та ефективність.

2. LTSM модель - є однією з найпоширеніших моделей для аналізу рядів динаміки. Вона продемонструвала задовільні результати і зуміла достатньо точно передбачити напрям та тенденцію руху кількості пацієнтів в лікарнях. LTSM модель зуміла визначити складні залежності в даних і показати задовільну точність прогнозування напряму тренду госпіталізованих пацієнтів.

3. Decision Forest модель - була використана для аналізу даних Covid-19. Ця модель, яка базується на ансамблі дерев рішень, показала загалом правильне розпізнавання тренду і напрямку руху кількості пацієнтів в лікарнях. Хоча, порівняно з іншими моделями, DecisionForest модель показала найменший рівень точності прогнозу.

В результаті отриманих результатів дослідження можна зазначити, що ARIMA модель показала найбільш точні результати прогнозування кількості пацієнтів в лікарнях. На основі цих даних, можно також відповісти на питання вибору найбільш оптимальної моделі для використання в інформаційному модулі, і обрати для цих цілей ARIMA модель, яка демонструє результати точності моделі на рівні 98 відсотків для прогнозування кількості пацієнтів в лікарнях.

ДЯКУЮ ЗА УВАГУ!