Міністерство освіти і науки України

Західноукраїнський національний університет

Факультет комп’ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

Звіт до лабораторної роботи №9

З дисципліни “Інтелектуальний аналіз даних”

Студента групи КН-31

Ковальковського В. В.

Тернопіль 2023

**Моделювання та прогнозування на основі часового ряду**

**Завдання:** обрати набір даних, який містить часовий ряд з сайту <https://www.ukrstat.gov.ua/>, та провести попередню обробку. Набір даних, має бути за період в 10 років з кроком 1 місяць та мати не менше трьох параметрів.

**Хід роботи**:

Для цієї лабораторної роботи я обрав набір даних про «Індекси споживчих цін на товари та послуги за 2013 – 2023 рр.». Це показники, що відображають зміни в рівні цін, які споживачі платять за корзину товарів та послуг протягом певного періоду часу. Вони використовуються для вимірювання інфляції, тобто зростання загального рівня цін, що в свою чергу впливає на купівельну спроможність грошей.

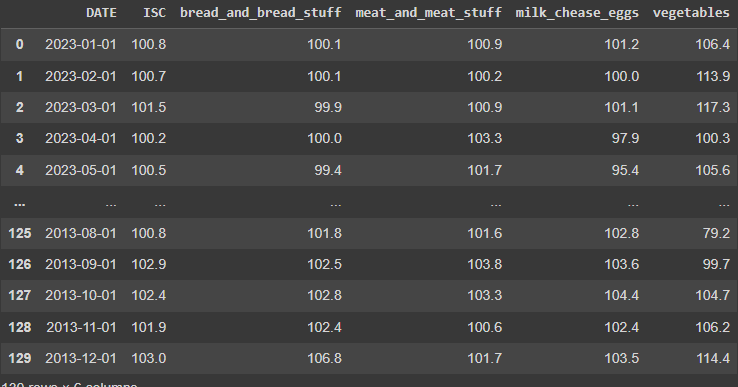


Рис.1 – Набір даних

На рис.1 показано датасет, який використовується у цій лабораторній роботі.

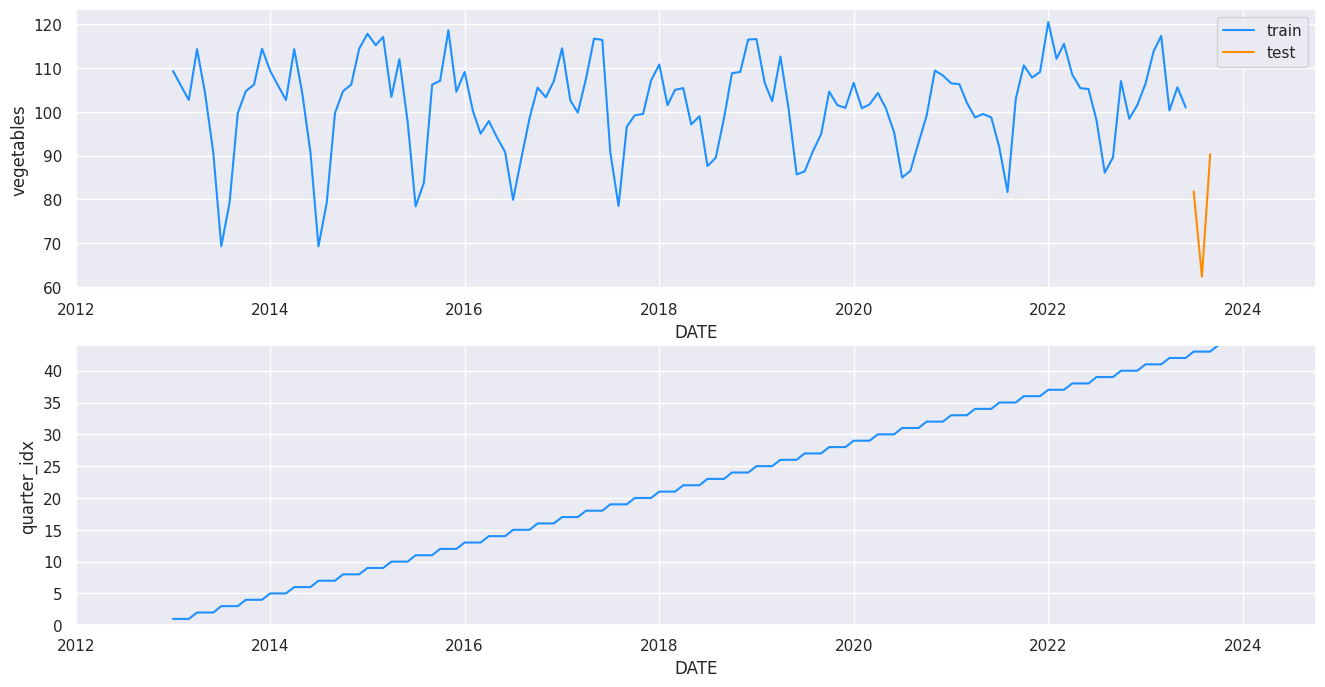


Рис. 2 – Графічне зображення набору даних та тестових даних.

**Наївний метод**

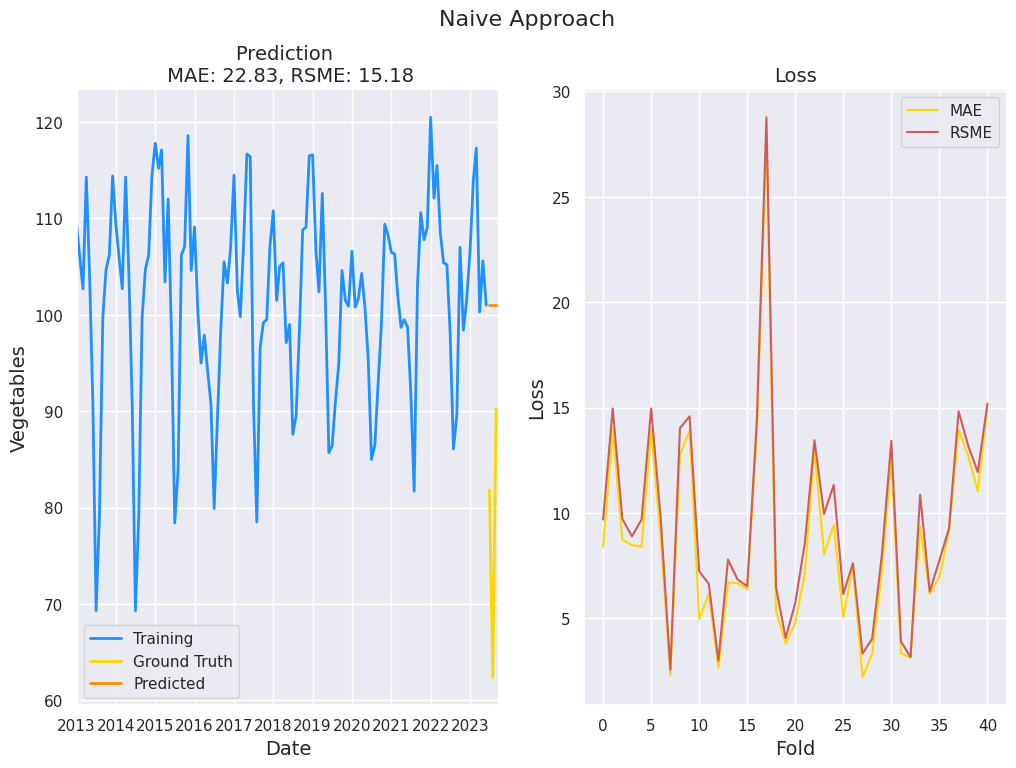
****

Рис. 3 – Графічне зображення наївного методу

На рисунку 3 показано два графіки, що представляють результати наївного методу прогнозування, але цього разу в контексті "Vegetables". Передбачувано, що це може бути дані щодо врожайності, цін або споживання овочів, але конкретний контекст не вказано.

Графік Прогнозування (Prediction):

На лівому графіку знову представлені три лінії: блакитна (Training), жовта (Ground Truth), і помаранчева (Predicted), які ілюструють тренувальні дані, фактичні дані та прогнозовані дані відповідно.

Схоже, що прогнозована лінія (Predicted) має деякі відповідності з фактичними даними (Ground Truth), але також є періоди, де прогнози значно відрізняються від фактичних даних, особливо на кінці періоду.

Метрики помилок залишаються незмінними: MAE - 22.83 і RSME - 15.18, що свідчить про середню точність моделі.

Графік Втрати (Loss):

На правому графіку показано динаміку помилок (MAE і RSME) протягом різних складів (Folds).

**Експоненціальне згладжування**

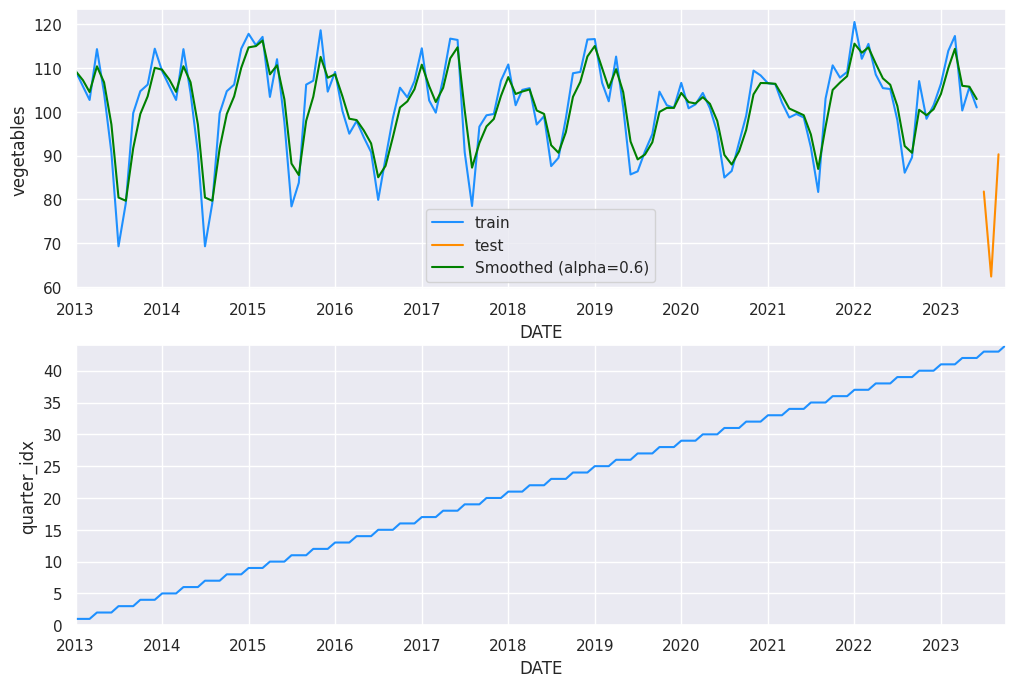


Рис. 4 – Графічне зображення методу експоненціального згладжування

На рисунка показано два графіки, що відображають результати застосування методу експоненційного згладжування для аналізу часових рядів, які, ймовірно, пов'язані з даними про овочі.

Верхній Графік (Дані про Овочі):

На графіку зображено три лінії: синя (train), яка представляє навчальний набір даних; оранжева (test), що представляє тестовий набір даних; і зелена (Smoothed), яка показує результати експоненційного згладжування з параметром гладкості (alpha=0.6).

Зелена лінія (Smoothed) виглядає гладшою порівняно з синьою та оранжевою лініями, що показує, як експоненційне згладжування може допомогти у зменшенні шуму в даних та виокремленні основної тенденції.

Нижній Графік (Індекс по Кварталах):

На другому графіку показано зростаючу лінію, яка, можливо, відображає кумулятивний індекс або підсумковий показник, пов'язаний з темою овочів, протягом кварталів з 2013 по 2023 рік.

За допомогою експоненційного згладжування можна ефективно оцінити та прогнозувати тенденції в даних, ігноруючи випадкові коливання, які не несуть значущої інформації. Цей метод особливо корисний, коли дані містять шум або коли прогнози потрібно швидко оновлювати з приходом нових даних.

**ARIMA**

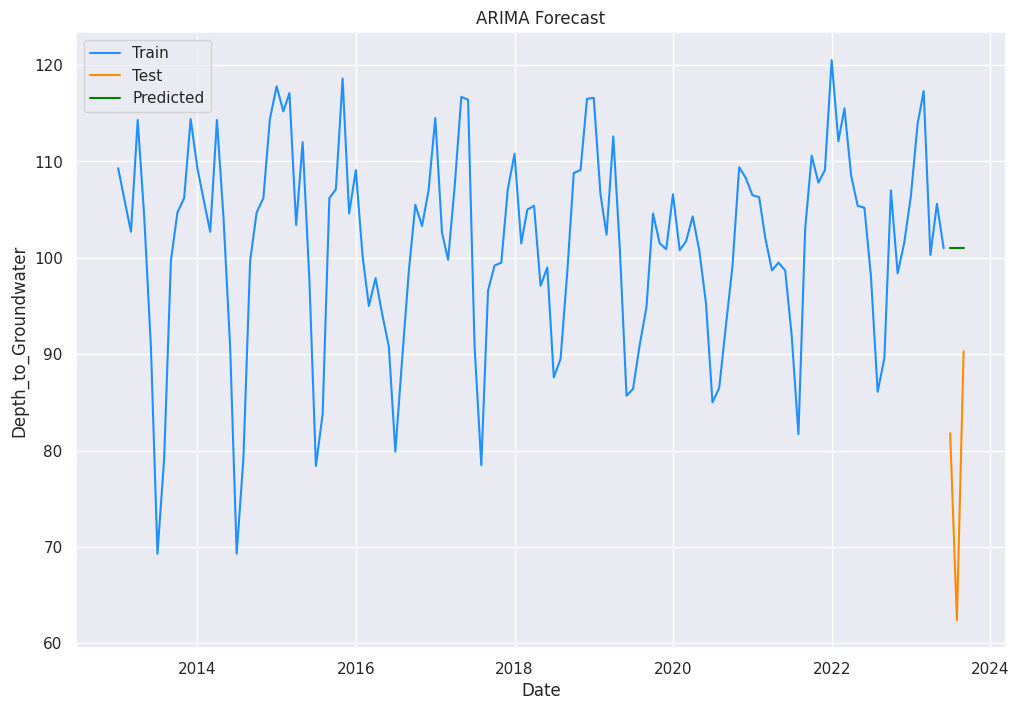


Рис. 5 – Графічне зображення методу ARIMA

Зображення показує графік прогнозування за допомогою методу ARIMA для часових рядів, які стосуються глибини до грунтових вод.

ARIMA, що стоїть за "Авторегресійна інтегрована модель ковзного середнього" (Autoregressive Integrated Moving Average), є популярним статистичним підходом до прогнозування часових рядів, який використовується для аналізу та прогнозування даних у часових рядах. Ось деякі ключові характеристики і компоненти методу ARIMA:

Авторегресія (AR): Ця компонента моделює залежність між спостереженням і деякою кількістю попередніх спостережень (званих лагами).

Інтегрованість (I): Цей аспект моделі вказує на кількість різниць, які необхідно використати для досягнення стаціонарності серії, тобто константності статистичних характеристик серії з часом.

Ковзне Середнє (MA): Ця частина моделі включає залежність між спостереженням і залишковою помилкою з моделі авторегресії, застосованою до попередніх спостережень.

На графіку представлені три різні лінії:

Синя лінія (Train): Відображає фактичні дані, які використовувалися для навчання моделі ARIMA.

Оранжева лінія (Test): Представляє фактичні дані, які використовувалися для перевірки точності прогнозів моделі.

Зелена лінія (Predicted): Це прогнозовані значення, отримані з моделі ARIMA.

З точки зору характеристики прогнозу:

Прогнозована лінія (Predicted) здається відстежує деякі з тенденцій і патернів, присутніх у тестовому наборі даних (Test), але з невеликими відхиленнями.

Проміжки, де прогнозовані дані не відповідають тестовим, можуть вказувати на потенційні області, де модель може бути покращена, наприклад, шляхом оптимізації параметрів ARIMA або включення сезонності, якщо така присутня.

**SARIMA**

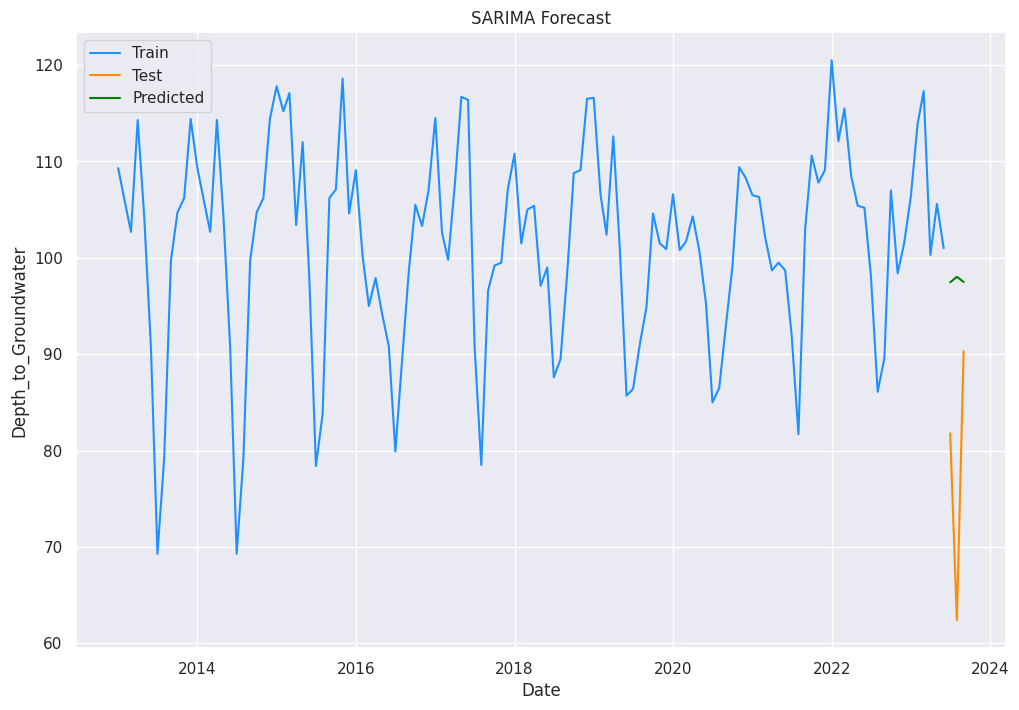


Рис. 6 – Графічне зображення методу SARIMA

**Порівняння MAE i RMSE для різних методів**

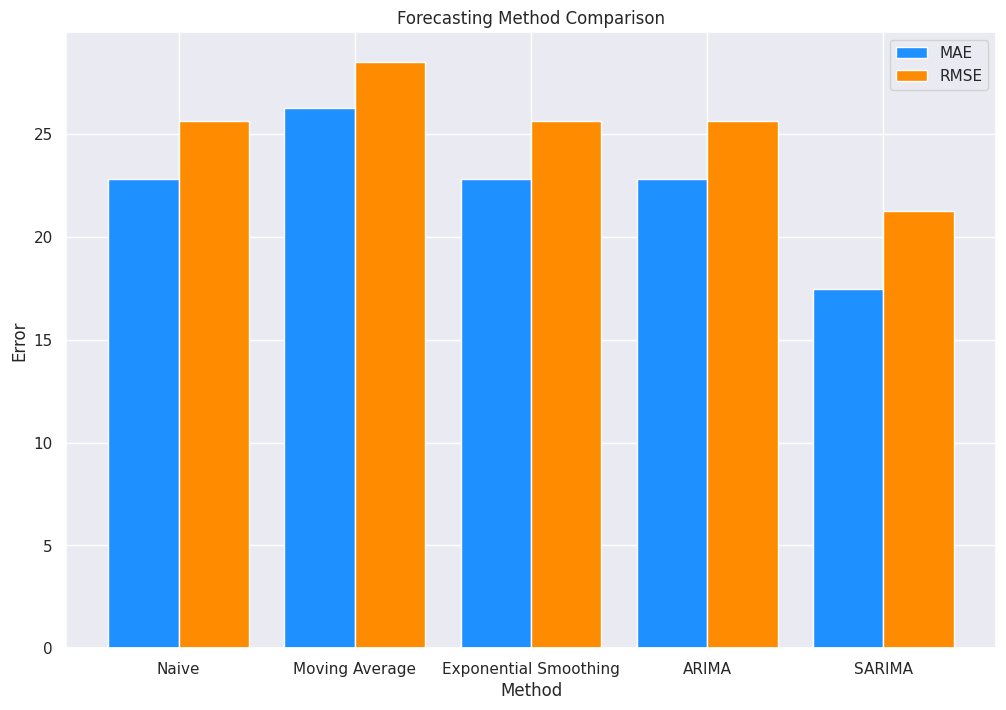
****

Рис. 7 – Графічне зображення порівняння методів

Method MAE RMSE

0 Naive 22.833333 25.645597

1 Moving Average 26.253125 28.534681

2 Exponential Smoothing 22.833333 25.645597

3 ARIMA 22.833333 25.645597

4 SARIMA 17.489403 21.259283

Кожен метод прогнозування представлений парою стовпчиків, де синій стовпчик відображає MAE, а оранжевий — RMSE. Висота стовпчиків відображає величину помилки для кожної метрики: чим вищий стовпчик, тим більша помилка. Важливо зазначити, що RMSE більш чутливий до викидів, оскільки він підносить помилки до квадрату перед усередненням, тому він часто вищий, якщо в даних присутні значні викиди.

На графіку зображено наступні методи прогнозування:

Наївний (Naive): Цей метод припускає, що наступне значення буде таким же, як і останнє спостереження.

Ковзне середнє (Moving Average): Прогнози засновані на середньому значенні попередніх спостережень.

Експоненційне згладжування (Exponential Smoothing Method): Вагове середнє попередніх спостережень з вагами, що експоненційно зменшуються.

ARIMA (Autoregressive Integrated Moving Average): Комбінує авторегресію, інтегрованість та ковзне середнє для моделювання часових рядів.

SARIMA (Seasonal ARIMA): Розширення ARIMA, що включає компоненти для моделювання сезонності в часових рядах.

З точки зору метрик помилок:

MAE вимірює середню абсолютну різницю між прогнозованими та фактичними значеннями, що дає загальне уявлення про помилки без урахування їх напрямку.

RMSE вимірює середній розмір помилки в тих самих одиницях, що і самі дані, і штрафує більші помилки більше, ніж менші.

Порівняння методів показує, що SARIMA має найменші помилки за обома метриками серед представлених методів, що робить його потенційно найкращим вибором для цих конкретних даних. Однак, вибір методу прогнозування залежить не тільки від помилок, але й від специфіки даних, вимог до моделі та контексту, у якому будуть використовуватися прогнози.

Висновок: під час роботи над лабораторною роботою, застосував різні методи прогнозування для набору даних, проаналізував результати виконання та визначив продуктивність кожного з методів.