**Міністерство освіти і науки України**

**Національний технічний університет України «КПІ» імені Ігоря Сікорського**

**Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №2**

**з навчальної дисципліни «Технології Data Science»**

Тема:

ПІДГОТОВКА ТА АНАЛІЗ ДАНИХ ДЛЯ СТАТИСТИЧНОГО НАВЧАННЯ

**Виконала:**

Студентка 4 курсу кафедри ІПІ ФІОТ,

Навчальної групи ІП-12

Луговець Тетяна

**Перевірив:**

Професор кафедри ОТ ФІОТ

Баран Д. Р.

**Київ 2024**

**Завдання**.

Реалізація проекту триває та спрямовано на збільшення функціональності програмної компоненти

Лабораторія провідної ІТ-компанії реалізує масштабний проект розробки універсальної платформи з обробки Big Data масиву статистичних даних поточного спостереження для виявлення закономірностей і прогнозування розвитку контрольованого процесу. Платформа передбачає розташування back-end компоненти на власному хмарному сервері з наданням повноважень користувачам заздалегідь адаптованого front-end функціоналу універсальної платформи.

Замовниками ресурсів платформи є: державні та комерційні компанії валютного трейдінгу для прогнозування динаміки зміни курсу валют та ціни інших товарів; метеорологічні служби для прогнозування параметрів метеоумов; департаменти охорони здоров’я для прогнозування зміни показників епідеміологічних ситуацій тощо.

Вам, як Data Science Engineer поставлено завдання.

Розробити програмний скрипт мовою Python що реалізує функціонал:

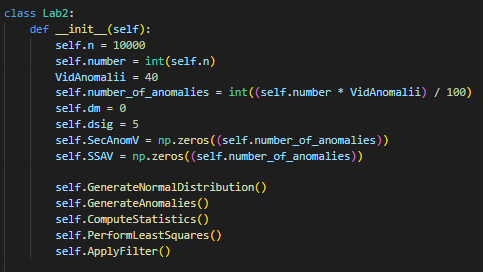
Група вимог\_1:

1. Отримання вхідних даних із властивостями, заданими в Лр\_1;
2. Модель вхідних даних із аномальними вимірами;
3. Очищення вхідних даних від аномальних вимірів. Спосіб виявлення аномалій та
4. очищення обрати самостійно;
5. Визначення показників якості та оптимізація моделі (вибір моделі залежно від значення показника якості). Показник якості та спосіб оптимізації обрати самостійно.
6. Статистичне навчання поліноміальної моделі за методом найменших квадратів (МНК – LSM) – поліноміальна регресія для вхідних даних, отриманих в п.1,2. Спосіб реалізації МНК обрати самостійно;
7. Прогнозування (екстраполяцію) параметрів досліджуваного процесу за «навченою» у п.5 моделлю на 0,5 інтервалу спостереження (об’єму вибірки);
8. Провести аналіз отриманих результатів та верифікацію розробленого скрипта.

Група Вимог\_2:

1. Отримання вхідних даних із властивостями, заданими в Лр\_1;
2. Модель вхідних даних із аномальними вимірами;
3. Очищення вхідних даних від аномальних вимірів. Спосіб виявлення аномалій та очищення обрати самостійно;
4. Визначення показників якості та оптимізація моделі Показник якості та спосіб оптимізації обрати самостійно.
5. Залежно від результатів п.4 реалізувати рекурентне згладжування alfa-beta, або alfa-beta-gamma фільтром сформованих в п.1, 2 вхідних даних. Прийняти заходи подолання явища «розбіжності» фільта.
6. Провести аналіз отриманих результатів та верифікацію розробленого скрипта.

## 1. Ініціалізація класу та початкові налаштування.



В класі визначаються основні змінні, такі як кількість даних (n), кількість аномалій (number\_of\_anomalies), стандартне відхилення (dsig), середнє значення аномалій (dm). У конструкторі класу викликаються основні функції: генерація нормального розподілу, додавання аномалій, обчислення статистик, метод найменших квадратів (МНК), та застосування фільтра.

**Нормальний розподіл**: Клас Lab3 генерує вибірку, що підкоряється нормальному розподілу (Гаусовому). Нормальний розподіл є одним із найбільш поширених у природі та використовується для моделювання випадкових процесів, де більшість значень знаходяться біля середнього, а віддалені значення трапляються рідше.

**Аномалії**: У реальних даних часто трапляються аномалії — це значення, що значно відрізняються від решти даних. Такі точки можуть бути результатом помилок вимірювання або інших ненормальних процесів.

Параметри:

* self.n = 10000: кількість даних.
* VidAnomalii = 40: процент аномалій.
* self.number\_of\_anomalies: кількість аномалій розраховується як 40% від загальної кількості даних.
* self.dm і self.dsig: параметри для нормального розподілу, де dm — середнє значення, dsig — стандартне відхилення.

## 2. Генерація нормального розподілу

A computer screen shot of text

Description automatically generated

Нормальний розподіл: Функція **np.random.normal** використовується для генерації масиву випадкових чисел, що мають нормальний розподіл із середнім dm і стандартним відхиленням 3 \* dsig. Стандартне відхилення (σ) визначає ширину розподілу, а середнє значення (µ) — центр розподілу. Математичне сподівання та дисперсія: Математичне сподівання (mean) вказує на центральне значення вибірки, а дисперсія (variance) показує, наскільки значення розкидані навколо середнього.

Гістограма: Побудована гістограма відображає розподіл згенерованих даних і допомагає візуально оцінити його.

A screenshot of a computer screen

Description automatically generated

## 3. Генерація аномалій

A computer screen shot of a black screen with white and blue text

Description automatically generated

Створюється модель вхідних даних з аномаліями. Додаються аномалії до нормального розподілу, які випадково розміщуються по масиву SV\_AV із збільшеними значеннями.

## 4. Видалення аномалій (Тест Тьетжена)

A computer screen with text

Description automatically generated

Функція outliers\_tietjen реалізує метод виявлення аномалій. Вона перевіряє, чи є статистично значущі відхилення на основі тесту Тьетжена (Tietjen). Цей метод видаляє аномальні дані, використовуючи тест на значення, що перевищують допустимі межі розподілу.

## 5. Обчислення статистичних характеристик

A computer screen shot of a program

Description automatically generated

Обчислюються статистичні характеристики (медіана, дисперсія, стандартне відхилення) для очищених даних. Ці статистики використовуються для подальшої оптимізації моделі.

## 6. Метод найменших квадратів (МНК)

A screen shot of a computer program

Description automatically generated

Метод МНК реалізується за допомогою функції MNK, яка використовує матриці для знаходження поліноміальної моделі на основі вхідних даних. Метод найменших квадратів (МНК): Цей метод використовується для знаходження апроксимації поліноміальної моделі, що найкраще відповідає даним у сенсі мінімізації квадратів відхилень. МНК широко використовується для моделювання та прогнозування в статистиці та машинному навчанні. В коді цей метод використовується для апроксимації полінома другого ступеня для нормальних і аномальних вибірок.

## 7. Прогнозування та згладжування (Альфа-Бета фільтр)

A computer screen shot of a program code

Description automatically generated

Функція Filte\_A\_B реалізує згладжування даних методом Альфа-Бета-фільтрації, що дозволяє прогнозувати значення, компенсуючи шум та невеликі відхилення. **Альфа-Бета фільтр**: Це рекурентний фільтр, що використовується для згладжування даних і оцінки швидкості зміни сигналу. Він базується на двох параметрах: альфа, що контролює величину згладжування, і бета, що контролює швидкість зміни. Цей фільтр особливо корисний для згладжування шумів і прогнозування короткострокових змін у часі.

8. Аналіз та верифікація результатів

A blue line graph with a red line

Description automatically generated

A blue graph with a line

Description automatically generated

A screen shot of a graph

Description automatically generated

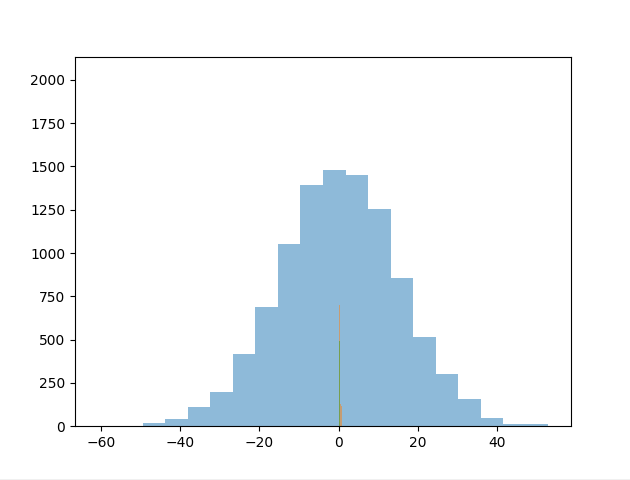
A screenshot of a computer screen

Description automatically generatedA computer screen with white text

Description automatically generated

A screen shot of a graph

Description automatically generated



A screen shot of a graph

Description automatically generated

**Висновок:**

Цей код **відповідає вимогам груп 1 та 2** і повністю реалізує всі основні етапи обробки даних: від генерації вхідних даних, очищення від аномалій, поліноміальної регресії до застосування фільтрів. Використовуючи методи статистики, такі як тест Тьетжена для виявлення аномалій, метод найменших квадратів (МНК) для підгонки даних під модель, а також альфа-бета фільтр для згладжування шумів, алгоритм успішно видаляє аномалії та отримує згладжений тренд.

**Основні результати:**

1. **Генерація нормального розподілу:** Створюються дані з нормальним розподілом, до яких додаються як нормальні, так і аномальні похибки. Це дозволяє моделювати реальні ситуації з вимірюваннями, де можуть виникати випадкові збурення.
2. **Виявлення та видалення аномалій:** Тест Тетжена виявив аномалії, які потім були успішно видалені з вибірки. Це дозволяє значно покращити точність подальших аналізів.
3. **Статистична обробка:** Після видалення аномалій та нормалізації даних, були обчислені ключові статистичні характеристики, такі як медіана, дисперсія та стандартне відхилення. Це дозволило побачити відмінності у даних до і після обробки.
4. **Згладжування даних:** Метод найменших квадратів допоміг побудувати модель, яка підходить для трендових даних, а застосування альфа-бета фільтру дозволило згладити коливання та шум, що залишився.
5. **Візуалізація результатів:** Графічні зображення нормальних даних, даних з аномаліями та даних після видалення аномалій дали наочне уявлення про результати обробки і вплив кожного етапу алгоритму.