

**Міністерство освіти і науки України Національний технічний університет України  
«КПІ» імені Ігоря Сікорського Кафедра інформатики та програмної інженерії ФІОТ**

**ЗВІТ з лабораторної роботи №1 з навчальної дисципліни «Технології Computer  
Vision»**

**Тема:**

**ПІДГОТОВКА ТА АНАЛІЗ ДАНИХ ДЛЯ СТАТИСТИЧНОГО НАВЧАННЯ**

**Виконав**

Студент 3 курсу кафедри ІІІ ФІОТ,  
Навчальної групи ІІІ-12  
Васильєв Є.К.

**Перевірив**

Професор кафедри ОТ ФІОТ  
Писарчук О.О.

**Київ 2023**

## **I. Мета:**

Підготувати та проаналізувати дані для статистичного навчання

## **II. Завдання:**

Завдання III рівня – максимально 9 балів.

1. Провести парсинг самостійно обраного сайту. Вміст даних, що підлягають парсингу – обрати самостійно.
2. Результати парсингу зберегти у файлі. Тип файлу обрати самостійно.
3. Оцінити динаміку тренду реальних даних.
4. Здійснити визначення статистичних характеристик результатів парсингу.
5. Синтезувати та верифікувати модель даних, аналогічних за трендом і статистичними характеристиками реальним даним, які є результатом парсингу.
6. Провести аналіз отриманих результатів.

## **III. Результати виконання лабораторної роботи.**

### **3.1. Синтезована математична модель;**

Після парсингу, оцінки динаміки тренду та визначення статистичних характеристик реальних даних, було вирішено синтезувати квадратичну модель з параметрами, які наблизять її статистичні показники до реальних даних.

Квадратична модель - це тип математичної моделі, яка використовується для апроксимації залежності між залежною змінною і незалежною змінною, де залежність між ними є квадратичною. У цій моделі враховується, що зміна в залежній змінній залежить від квадратичних змін у незалежній змінній.

Загальний вигляд квадратичної моделі має наступний вигляд:  $Y = a * X^2 + b * X + c$ , Де  $Y$  – залежна змінна, яку ми намагаємось передбачити,  $X$  – незалежна змінна, яка впливає на  $Y$ ,  $a, b, i c$  – коефіцієнти моделі, які потрібно оцінити.

Квадратична модель має три параметри:  $a, b, i c$ . Параметр  $a$  відповідає за ступінь квадратичного впливу  $X$  на  $Y$ . Якщо  $a$  додатне, то це означає, що зміна  $X$  спричиняє квадратичний ріст в  $Y$ . Якщо  $a$  від'ємне, то зміна  $X$  спричиняє квадратичне зменшення в  $Y$ .

Параметр  $b$  представляє лінійний вплив  $X$  на  $Y$ . Він вказує на те, як зміна  $X$  впливає на  $Y$  в лінійному режимі.

Параметр  $c$  - це константа, або відома як вільний член, який вказує на значення  $Y$ , коли  $X$  рівний нулю.

### **3.2. Результати архітектурного проектування та їх опис;**

Після аналізу характеру поведінки досліджуваних даних, та підбору оптимальних параметрів, було визначено наступні коефіцієнти для квадратичної моделі:

$a = 12, b = -850, c = 30000$ . Для оцінки якості моделі, окрім візуальної перевірки, було також обраховано відсоткову різницю між середнім значенням, медіаною, дисперсією та середнім квадратичним відхиленням реальних даних і квадратичною регресією.

### 3.3. Опис структури проекту програми;

Структура проекту програми наведена на блок-схемі:

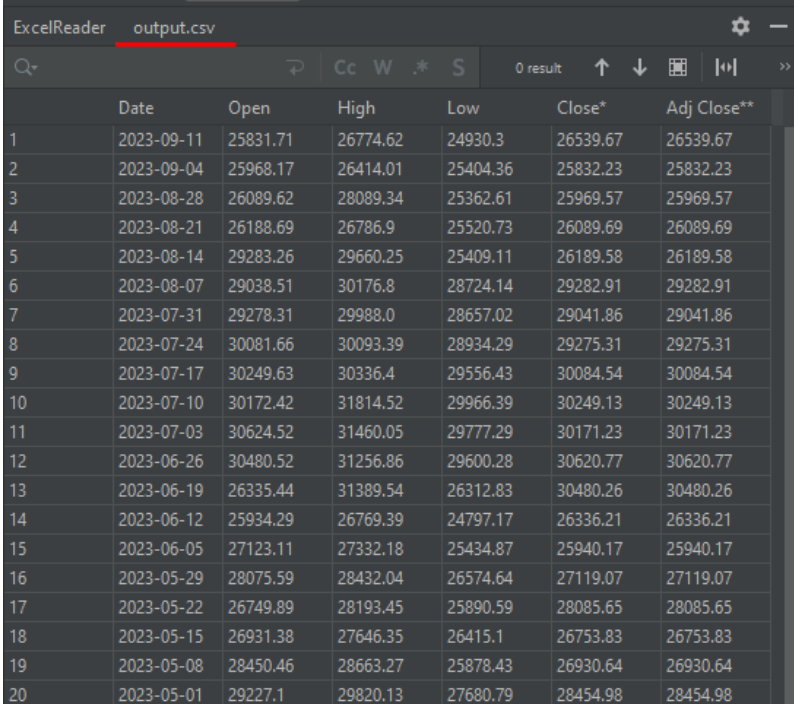


Рис.1. Блок схема алгоритму програми.

Робота алгоритму починається з отримання даних про курс біткоїна за останні 100 тижнів шляхом парсингу сайту [finance.yahoo.com](https://finance.yahoo.com). Отримані дані очищуються та трансформуються для зручності подальшої роботи з ними, також вони зберігаються у .csv файл. Після цього, для оцінки характеру зміни даних, вони візуалізуються з використанням лінійного графіку. Далі визначається математичне сподівання, медіана, дисперсія та середньо квадратичне відхилення даних. Потім підбираються параметри для квадратичної регресії та створюється відповідна модель, яка одразу додається на графік разом з початковими даними. Останнім кроком визначаються статистичні характеристики синтезованої моделі та порівнюються з реальними даними.

### 3.4. Результати роботи програми відповідно до завдання;

Результати роботи програми наведені на відповідних скріншотах:



	Date	Open	High	Low	Close*	Adj Close**
1	2023-09-11	25831.71	26774.62	24930.3	26539.67	26539.67
2	2023-09-04	25968.17	26414.01	25404.36	25832.23	25832.23
3	2023-08-28	26089.62	28089.34	25362.61	25969.57	25969.57
4	2023-08-21	26188.69	26786.9	25520.73	26089.69	26089.69
5	2023-08-14	29283.26	29660.25	25409.11	26189.58	26189.58
6	2023-08-07	29038.51	30176.8	28724.14	29282.91	29282.91
7	2023-07-31	29278.31	29988.0	28657.02	29041.86	29041.86
8	2023-07-24	30081.66	30093.39	28934.29	29275.31	29275.31
9	2023-07-17	30249.63	30336.4	29556.43	30084.54	30084.54
10	2023-07-10	30172.42	31814.52	29966.39	30249.13	30249.13
11	2023-07-03	30624.52	31460.05	29777.29	30171.23	30171.23
12	2023-06-26	30480.52	31256.86	29600.28	30620.77	30620.77
13	2023-06-19	26335.44	31389.54	26312.83	30480.26	30480.26
14	2023-06-12	25934.29	26769.39	24797.17	26336.21	26336.21
15	2023-06-05	27123.11	27332.18	25434.87	25940.17	25940.17
16	2023-05-29	28075.59	28432.04	26574.64	27119.07	27119.07
17	2023-05-22	26749.89	28193.45	25890.59	28085.65	28085.65
18	2023-05-15	26931.38	27646.35	26415.1	26753.83	26753.83
19	2023-05-08	28450.46	28663.27	25878.43	26930.64	26930.64
20	2023-05-01	29227.1	29820.13	27680.79	28454.98	28454.98

Рис.2. Очищені та перетворені дані з сайту записані у файл.

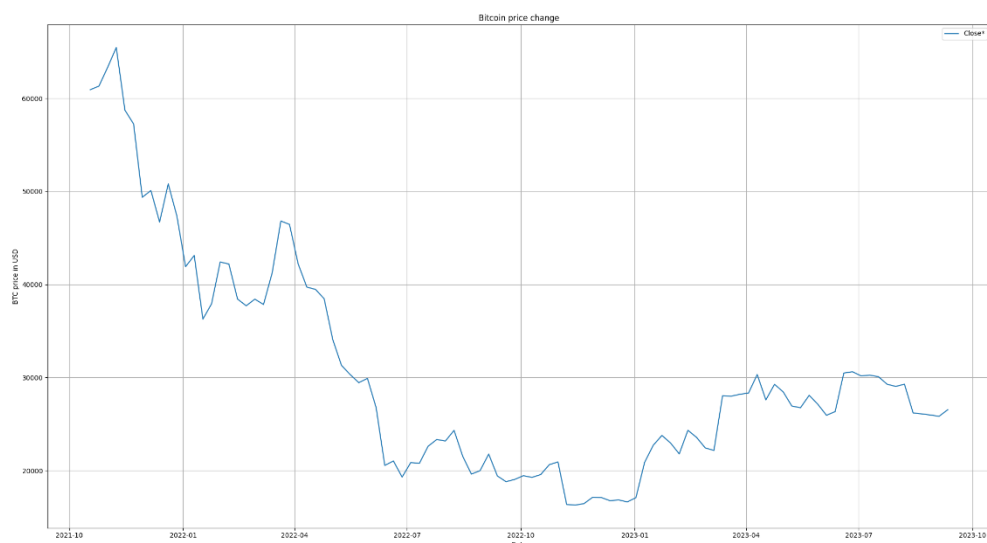


Рис.3. Візуалізація даних.

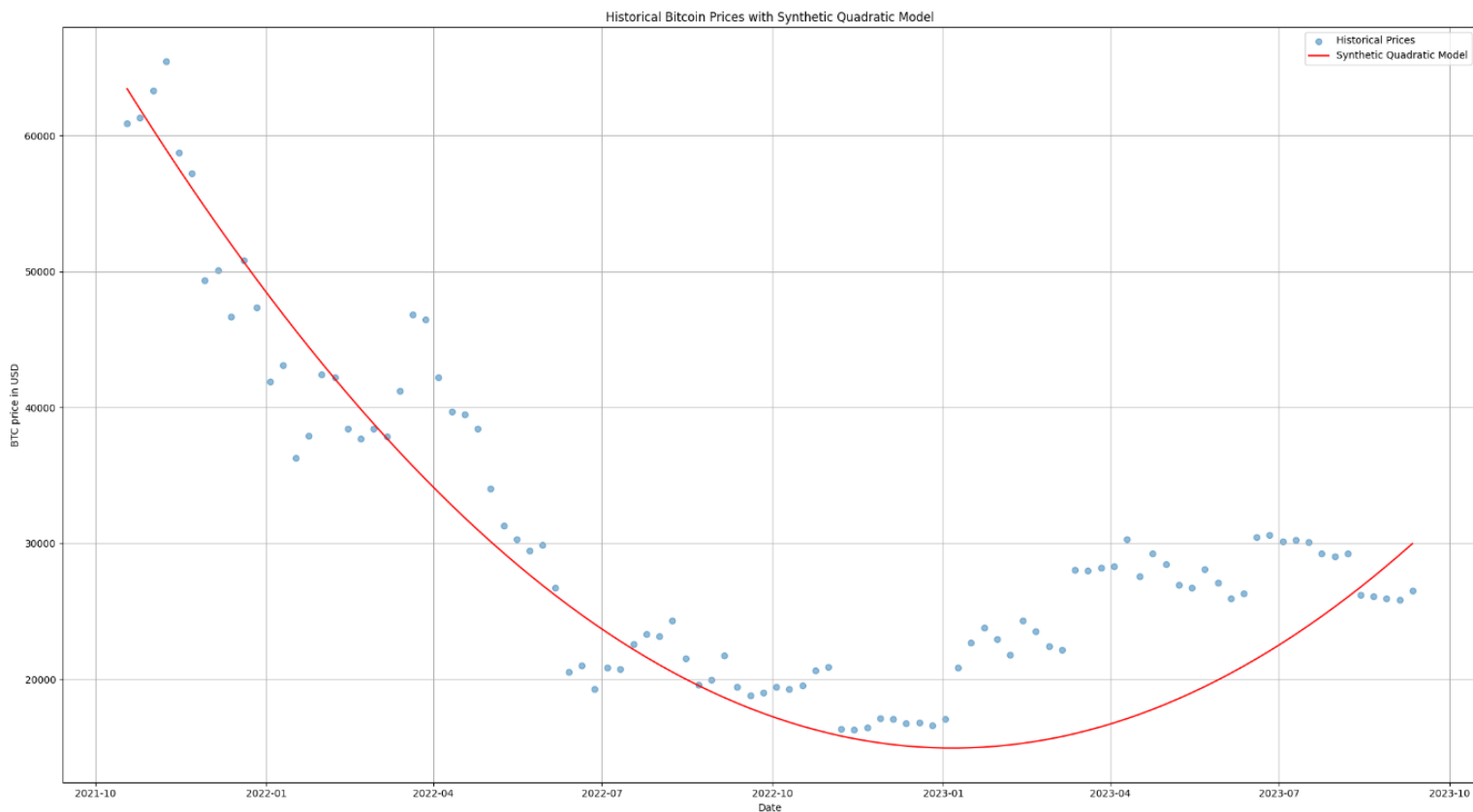


Рис.4. Візуалізація синтезованої моделі.

```

Statistical characteristics for the dataset
Mean: 30341.92
Median: 27355.22
Variance: 138221042.10
Root mean square deviation: 11756.74

Statistical characteristics for the synthetic model
Mean: 27327.00
Median: 22450.00
Variance: 175153816.20
Root mean square deviation: 13234.57

Difference between real data and synthetic model:
Mean: -9.94%
Median: -17.93%
Variance: 26.72%
Root mean square deviation: 12.57%

```

Рис.5. Статистичні характеристики початкових даних, створеної моделі, та їх порівняння.

### 3.5. Программний код, що забезпечує отримання результату.

```
'''
=====
Виконав: Васильєв Єгор
Lab_work_1, III рівень складності
'''
=====

import requests
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from bs4 import BeautifulSoup as bs

# constants for accessing and parsing finance.yahoo.com
url = "https://finance.yahoo.com/quote/BTC-USD/history?period1=1631577600&period2=1694649600&interval=1wk&filter=history&frequency=1wk&includeAdjustedClose=true"
user_agent = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/104.0.5112.79 Safari/537.36"
headers = {"User-Agent": user_agent}

2 usages
def stat_characteristics(lst):
    '''
    =====
    :param lst: sample for which characteristics are calculated
    :return: mean, median, variance, and root-mean-square deviation of the sample
    '''
    =====
    mean = np.mean(lst)
    median = np.median(lst)
    variance = np.var(lst)
    RMSD = np.sqrt(variance)
    return mean, median, variance, RMSD

2 usages
def print_characteristics(mean, median, variance, RMSD):
    '''
    =====
    :param mean: mean
    :param median: median
    :param variance: variance
    :param RMSD: root-mean-square deviation
    :return: void
    '''
    =====
    print(f"Mean: {mean:.2f}")
    print(f"Median: {median:.2f}")
    print(f"Variance: {variance:.2f}")
    print(f"Root mean square deviation: {RMSD:.2f}\n")

4 usages
def percentage_difference(old_value, new_value):
    '''
    =====
    :param old_value: first number
    :param new_value: second number
    :return: percentage difference between two numbers
    '''
    =====
    return ((new_value - old_value) / old_value) * 100
```

```

# parsing the website
response = requests.get(url, headers=headers)
soup = bs(response.content, features="html.parser")
temperatures_html = soup.find_all(name="table", class_="W(100%) M(0)")[0]

# retrieving data from the html table
btc_rate = []

for row in temperatures_html.find_all("tr"):
    row_data = []

    for cell in row.find_all("td"):
        row_data.append(cell.text)

    btc_rate.append(row_data)

columns = temperatures_html.find_all("th")
column_list = [item.get_text(strip=True) for item in columns]
rates = pd.DataFrame(btc_rate, columns=column_list)

# data cleansing and transformation
rates = rates.dropna()
rates['Open'] = rates['Open'].str.replace(',', '').astype(float)
rates['High'] = rates['High'].str.replace(',', '').astype(float)
rates['Low'] = rates['Low'].str.replace(',', '').astype(float)
rates['Close*'] = rates['Close*'].str.replace(',', '').astype(float)
rates['Adj Close**'] = rates['Adj Close**'].str.replace(',', '').astype(float)
rates['Volume'] = rates['Volume'].str.replace(',', '').astype(float)
rates['Date'] = pd.to_datetime(rates['Date'], format='%b %d, %Y').dt.date
rates.to_csv("output.csv")

# data visualisation
rates.plot(figsize=(8, 6), x='Date', y='Close*')
plt.grid(True)
plt.ylabel('BTC price in USD')
plt.title('Bitcoin price change')
plt.show()

# calculating statistical characteristics for the data
print("Statistical characteristics for the dataset")
mean_data, median_data, variance_data, RMSD_data = stat_characteristics(rates['Close*'])
print_characteristics(mean_data, median_data, variance_data, RMSD_data)

```

```

# synthesis of a model similar to real data
a = 12
b = -850
c = 30000
x_values_synthetic = [x for x in range(len(rates))]
y_values_synthetic = [a * x ** 2 + b * x + c for x in x_values_synthetic]

# visualisation of quadratic model along with real data
plt.figure(figsize=(12, 6))
plt.scatter(rates['Date'], rates['Close*'], label='Historical Prices', alpha=0.5)
plt.plot(*args: rates['Date'], y_values_synthetic, color='red', label='Synthetic Quadratic Model')
plt.xlabel('Date')
plt.ylabel('BTC price in USD')
plt.title('Historical Bitcoin Prices with Synthetic Quadratic Model')
plt.legend()
plt.grid(True)
plt.show()

# calculating statistical characteristics for the synthetic model
print("Statistical characteristics for the synthetic model")
mean_synthetic, median_synthetic, variance_synthetic, RMSD_synthetic = stat_characteristics(y_values_synthetic)
print_characteristics(mean_synthetic, median_synthetic, variance_synthetic, RMSD_synthetic)

# calculating difference between statistical characteristics of real data and artificial model
print("Difference between real data and synthetic model:")
print(f"Mean: {percentage_difference(mean_data, mean_synthetic):.2f}%")
print(f"Median: {percentage_difference(median_data, median_synthetic):.2f}%")
print(f"Variance: {percentage_difference(variance_data, variance_synthetic):.2f}%")
print(f"Root mean square deviation: {percentage_difference(RMSD_data, RMSD_synthetic):.2f}%")

```

#### IV. Висновки.

В ході лабораторної роботи було здійснено парсинг даних курсу біткоїна до долара за останні 100 тижнів і досліджено характер цих даних. Відповідно до них було створено та проаналізовано квадратичну модель. Незважаючи на те, що статистичні характеристики отриманої регресії не сильно відрізняються від початкових даних, такий спосіб створення моделей не є ефективним. Доцільніше для прогнозування даних використовувати автоматичні методи навчання математичних моделей. Загалом, було вивчено усі етапи підготовки та аналізу даних для статистичного навчання.

Виконав: студент Васильєв Єгор