

**Міністерство освіти і науки України**  
**Львівський національний університет імені Івана Франка**

Факультет електроніки та комп'ютерних технологій

**Звіт**

Про виконання лабораторної роботи №4

**3 курсу** «Використання GPT-4 для статистичного аналізу даних »

Виконав:

Студент групи Фес-21

Шавало Андрій

Львів-2025

**Мета роботи:** Ознайомитися з можливостями використання мовних моделей (GPT-4 або аналогів) для автоматизації статистичного аналізу даних. Навчитися інтегрувати OpenAI API в Python для проведення описової статистики, перевірки розподілу та пошуку залежностей у вибірці.

### Завдання

1. Завантаження та підготовка даних. Оберіть або згенеруйте реальний чи штучний датасет (наприклад, Kaggle, sklearn.datasets, pandas.DataFrame). Виконайте базову предобробку: заповнення пропущених значень, видалення дублікатів, нормалізацію даних.

```
df = pd.read_csv('2018.csv')
```

2. Описова статистика. Використовуючи pandas, обчисліть основні статистичні параметри (середнє, медіана, дисперсія, стандартне відхилення, мін/макс значення). Передайте цей аналіз в GPT-4 через API та попросіть модель зробити висновки щодо даних.

```
data_overview = {
    'Інформація про дані': str(df.info()),
    'Опис даних': str(df.describe()),
    'Кількість пропущених значень': str(df.isnull().sum())
}
```

```
response = client.chat.completions.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "system", "content": "Ви - аналітик даних, який пояснює"},
        {"role": "user", "content": f"Поясніть ці результати: {report}"}
    ]
)
```

3.Перевірка нормальності розподілу Обчисліть асиметрію (skewness) та ексцес (kurtosis) для числових змінних. Виконайте статистичні тести (Шاپіро-Уїлка, Колмогорова-Смірнова) за допомогою scipy.stats. Надішліть результати тестів у GPT-4 та отримайте пояснення щодо нормальності розподілу.

```
correlation_matrix = df.corr(numeric_only=True)

skewness = df.select_dtypes(include=np.number).apply(lambda x: stats.skew(x))
kurtosis = df.select_dtypes(include=np.number).apply(lambda x: stats.kurtosis(x))

shapiro_results = {}
for column in df.select_dtypes(include=np.number).columns:
    shapiro_results[column] = stats.shapiro(df[column])

ks_results = {}
for column in df.select_dtypes(include=np.number).columns:
    ks_results[column] = stats.kstest(df[column], cdf='norm')
```

```
report = f"""
Звіт аналізу набору даних

1. Огляд даних:
Інформація про дані:
{data_overview['Інформація про дані']}

Опис даних:
{data_overview['Опис даних']}

Кількість пропущених значень:
{data_overview['Кількість пропущених значень']}

2. Кореляційна матриця:
{correlation_matrix.to_string()}

3. Тест Шاپіро-Уїлка:
{shapiro_results}
"""
```

4.Аналіз кореляції Обчисліть кореляцію між числовими змінними (df.corr()). Передайте отриману кореляційну матрицю моделі GPT-4 та попросіть її сформулювати текстове пояснення виявлених зв'язків.

```
correlation_matrix = df.corr(numeric_only=True)
```

5. Автоматичний генератор звіту Створіть Python-скрипт, який генерує повний аналіз датасету, передає результати в GPT-4 та отримує пояснення у вигляді зрозумілого тексту. Збережіть згенерований звіт у форматі .txt

```
try:
    response = client.chat.completions.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": "Ви - аналітик даних, який пояснює"},
            {"role": "user", "content": f"Поясніть ці результати: {report}"}
        ]
    )
    explanation = response.choices[0].message.content
except Exception as e:
    explanation = f"Помилка: {str(e)}"

with open('data_analysis_report.txt1', 'w', encoding='utf-8') as f:
    f.write('\n\nПояснення OpenAI українською:\n' + explanation)

print('Звіт згенеровано та збережено у файлі "data_analysis_report.txt"')
```

**Висновок:** Під час виконання цієї лабораторної роботи я ознайомився з можливостями використання мовних моделей, зокрема GPT-4, для автоматизації статистичного аналізу даних. Метою роботи було навчитися інтегрувати OpenAI API у Python для проведення описової статистики, перевірки розподілу даних та пошуку залежностей між змінними.

На першому етапі я завантажив датасет та виконав базову предобробку даних, включаючи заповнення пропущених значень, видалення дублікатів та нормалізацію. Це дозволило мені підготувати дані для подальшого аналізу.

Далі я провів описову статистику, обчисливши основні параметри, такі як середнє значення, медіана, дисперсія, стандартне відхилення, мінімальні та максимальні значення. Ці результати я передав у GPT-4 через API, щоб отримати інтерпретацію та висновки щодо характеристик даних.

Для перевірки нормальності розподілу я обчислив асиметрію та ексцес, а також виконав статистичні тести Шапіро-Вілка та Колмогорова-Смірнова. Отримані результати я також передав у GPT-4, щоб отримати пояснення щодо того, наскільки розподіл даних відповідає нормальному.

На завершальному етапі я провів аналіз кореляції між числовими змінними, обчисливши кореляційну матрицю. Цю матрицю я також

передав у GPT-4, щоб отримати текстове пояснення виявлених зв'язків між змінними.

Підсумовуючи, я створив автоматизований Python-скрипт, який виконує повний аналіз датасету, передає результати в GPT-4 та генерує зрозумілий текстовий звіт. Ця робота дозволила мені на практиці переконатися у потужності інтеграції мовних моделей для аналізу даних, що значно спрощує інтерпретацію результатів та формування висновків.

У результаті я отримав автоматично сформований звіт, який містить основні характеристики вибірки, висновки про нормальність розподілу та опис залежностей між змінними. Це дозволяє ефективно аналізувати дані та робити обґрунтовані висновки на основі отриманих результатів.