ПО 01. Аналіз даних
Лабораторна робота №0: Вибір даних та формування дослідницьких питань

Данило Тавров

1 Ідея роботи

Протягом усього семестру студенти будуть виконувати різні види аналізу даних (розвідковий, статистичний, регресійний, непараметричний, беєсівський, причиново-наслідковий тощо), намагаючись дати відповідь на цікаві дослідницькі питання з використанням **одного й того ж набору даних**.

В ідеалі передбачається, що в процесі виконання різних видів аналізу уточнюватимуться наявні та формуватимуться нові дослідницькі питання. Але на самому початку семестру від кожної команди дослідників очікується приблизний перелік дослідницьких питань, на які в принципі було б цікаво дістати відповіді, працюючи з відповідним набором даних.

2 Звітність щодо результатів роботи

Студенти повинні надати викладачу:

- інформацію щодо набору даних, який буде використано протягом семестру, де та за допомогою чого його буде добуто;
- перелік дослідницьких питань.

Набір даних може стосуватися **довільної** предметної області, із якою цікаво працювати команді. Це повинен бути **відносно великий** набір даних. Зокрема, він повинен мати такі характеристики:

- щонайменше **сотня тисяч** спостережень (рядків);
- щонайменше **декілька десятків** змінних (стовпців). Змінні повинні бути **різної** природи: і числові, і бінарні, і категорійні.

Деякі рекомендації щодо дослідницьких питань наведено в лекції. Варто **особливо наголосити**, що серед цих питань повинні бути питання причиново-наслідкового характеру. Доведення причиново-наслідкового характеру зв'язку між різними змінними є надзвичайно непростою задачею, і в нашому курсі ми зможемо розглянути тільки найпростіші підходи. Але набір даних повинен давати можливість у принципі ці підходи застосовувати. Отже потрібно перевіряти, що:

- причиново-наслідкові питання можна сформулювати;
- що в наборі даних наявна **достатня кількість** (біля десятка, а то й більше) змінних, які може бути враховано під час такого причиново-наслідкового аналізу.

Корисними джерелами даних є, зокрема, такі:

- **море цікавих даних**, на основі яких публікують наукові статті в найліпших журналах, зібрано на сайті openICPSR;
- Kaggle;

- дані Світового банку;
- відкриті урядові дані США;
- відкриті урядові дані України;
- IPUMS.