**Міністерство Освіти І НАУКИ України**

**Національний університет "Львівська політехніка"**

Інститут **КНІТ**

Кафедра **ПЗ**

### ЗВІТ

До лабораторної роботи № 1

**З дисципліни:** *“Видобування та опрацювання даних”*

**На тему:** *“**Розуміння задачі і області застосування обраного набору даних”*

**Лектор:**

доц. каф. ПЗ

Матвіїв Б. Д.

**Виконав:**

ст. гр. ПЗ-43

Солоненко О. О.

**Прийняв:**

асист. каф. ПЗ

Угриновський Б. В.

« \_\_\_\_ » \_\_\_\_\_\_\_\_ 2021 р.

∑= \_\_\_\_\_ .

**Тема роботи:** Розуміння задачі і області застосування обраного набору даних.

**Мета роботи:** Ознайомитись із поняттями задачі і області застосування обраного набору даних.

**ЗАВДАННЯ**

Вибір та обгрунтування набору даних. Фаза Business understanding за СRISP-DM.

Перша лабораторна робота призначена для обрання та мотивації обраного набору даних. Основними вимогами є:

* Табличне представлення даних (не зображення або часовий ряд)
* Мінімум 250 записів - чим краще тим більше , в рамках розумного (менша кількість записів можлива але обговорювана)
* Мінімум 4 колонки з даними по яким можна буде виконувати аналіз / моделювання і візуалізацію (колонки типу ID, номер або службові не рахуються)
* Один варіант набору даних на групу

**ХІД ВИКОНАННЯ**

1. Який набір даних ви обрали ?

[Airline Passenger Satisfaction](https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction) - дані зібрані з опитування про задоволеність послугами авіакомпанії

1. Які у вас ідеї / теорії / припущення що у вибраному домені можна покращити з аналізом та видобуванням даних ?

Для покращення і полегшення видобування відгуків про авіаперельоти можна створити автоматизовний процес збору відгуків про політ після його завершення шляхом проходження онлайн-опитування через додаток і/або електронну пошту

1. Які проблеми є у області з якої ви взяли дані ? Що є "bottleneck" у даному напрямі, що можна автоматизувати ?

Власне проблема задоволення якістю авіаперельотів. За допомогою аналізу та опрацювання даного датасету можна отримати певні критерії, які найбільше впливають на оцінку та задоволення пасажирів авіакомпанії. Тому при знаходженні найбільш вразливих точок авіакомпанія зможе знайти місця для покращення власного сервісу.

1. Що за ризики / складності ви припускаєте можуть очікувати вас в процесі роботи з даними ?

Загалом наразі я не бачу особливих складностей, які можуть мене очікувати в процесі роботи з даними. Одним з моїх варіантів може бути: неможливість визначити конкретні властивості, які впливають на оцінку сервісу авіаперельотів.

1. В рамках добування даних яким методом ваша проблема або ідея може бути вирішена? (кластеризація, прогнозування, класифікація)

Для даного датасету найкращим методом буде класифікація. Обраний датасет містить розбивку на тренувальний та тестовий датасет. Тренувальний містить набір властивостей та окреме поле, яке вказує на те, чи був цей пасажир задоволений своїм перельотом. Тестовий набір містить лише набір властивостей за допомогою яких ми маємо визначити чи був задоволений цей пасажир своїм польотом.

**ВИСНОВКИ**

На цій лабораторній роботі я ознайомитись із поняттями задачі і області застосування обраного набору даних. Обрав датасет для виконання подальших лабораторних робіт – датасет задоволення перельотами пасажирів авіакомпанії. Також проаналізував даний датасет, виявив шляхи для покращення та автоматизації процесу видобування даних. Описав основну проблему характерну для даного домену, ризики з якими можна зіткнутись при використанні даного датасету. А також описав метод обробки даних, який найбільше підійде для роботи з цим датасетом.