Прізвище: **КИРИЛЮК**
Ім'я: **Дмитро**
Група: **ПП-22**
Варіант: **08**
Дата захисту: **17.03.2025р**.

Кафедра: **САПР**
Дисципліна: **Системи інтелектуального аналізу та візуалізації даних**
Перевірив: **Андрій КЕРНИЦЬКИЙ**

ЗВІТ

до лабораторної роботи №04
на тему "**Ознайомлення з WEKA. Підготовка даних.**"

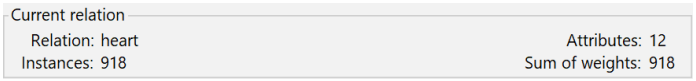
Мета роботи: ознайомлення студентів з системою WEKA, яка є потужним інструментом для обробки і аналізу даних. Студенти повинні навчитися використовувати основні функції цієї системи, зокрема, завантажувати, обробляти і візуалізувати набори даних. Додатково, метою є вміння проводити попередній аналіз даних і коректно вибирати методи їх обробки в майбутньому. Студенти мають розвинути вміння використовувати WEKA для практичного застосування у процесі вивчення курсу та роботи над індивідуальними завданнями. Результатом виконання роботи є підготований набір даних до подальшого аналізу та машинного навчання.

Індивідуальне завдання:
1. Визначте та охарактеризуйте набір даних.
2. Дослідження та попередня обробка даних.
3. Дослідить можливості Weka.

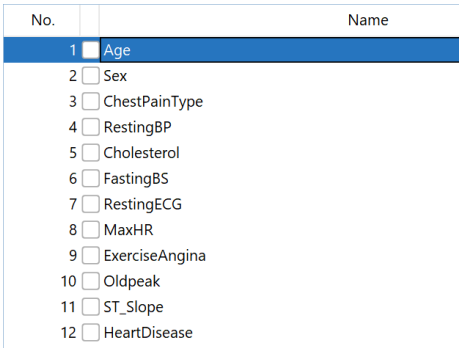
Результати виконання програми:

1. Вибраний датасет: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/

Практичним завданням для вирішення є знайти взаємозв'язок між медичними показниками пацієнта, щоб передбачити, чи можлива у нього серцева хвороба.



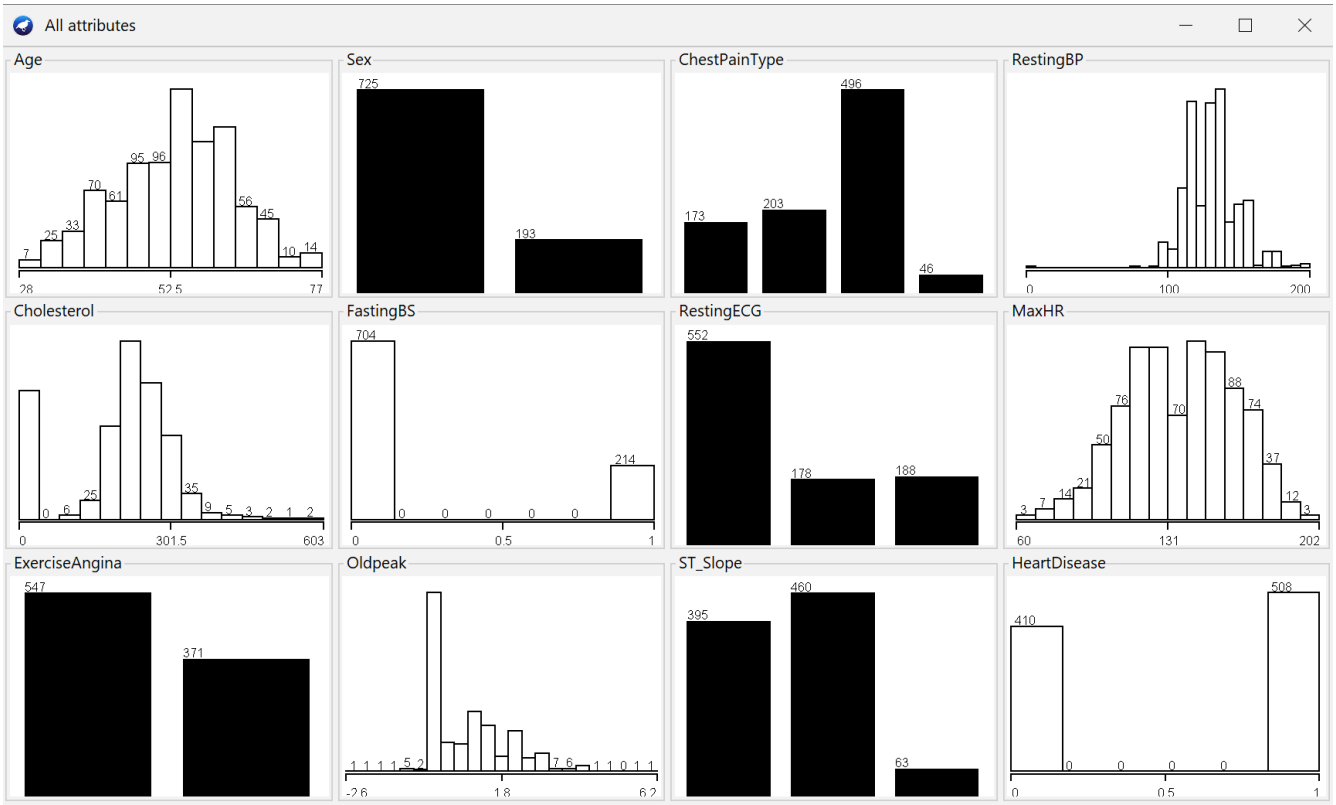
У вибірці 918 примірників.



Перелік атрибутів:

Age – вік
Sex – стать
ChestPainType – тип болю у грудях
RestingBP – артеріальний тиск у спокої
Cholesteron – сироватковий холестерин
FastingBs – рівень цукру в крові натще
RestingECG – результати електрокардіограми в спокої
MaxHR – досягнута максимальна частота серцевих скорочень
ExerciseAngina – стенокардія фізичного навантаження
Oldpeak – зниження сегмента ST на електрокардіограмі під час фізичного навантаження
ST_Slope – нахил піку навантаження на сегмент ST
HeartDisease – виявлена хвороба серця

Екземплярів із відсутніми значеннями немає.



Бачимо наявність викидів у атрибутах RestingBP, Cholesteron, MaxHR і Oldpeak відповідно.

Цільовим атрибутом є HeartDisease, який приймає значення 0 і 1 (нехворий і хворий на серце відповідно).

Клас хворих містить 410 екземплярів, а здорових – 506 екземплярів.

2.

a.

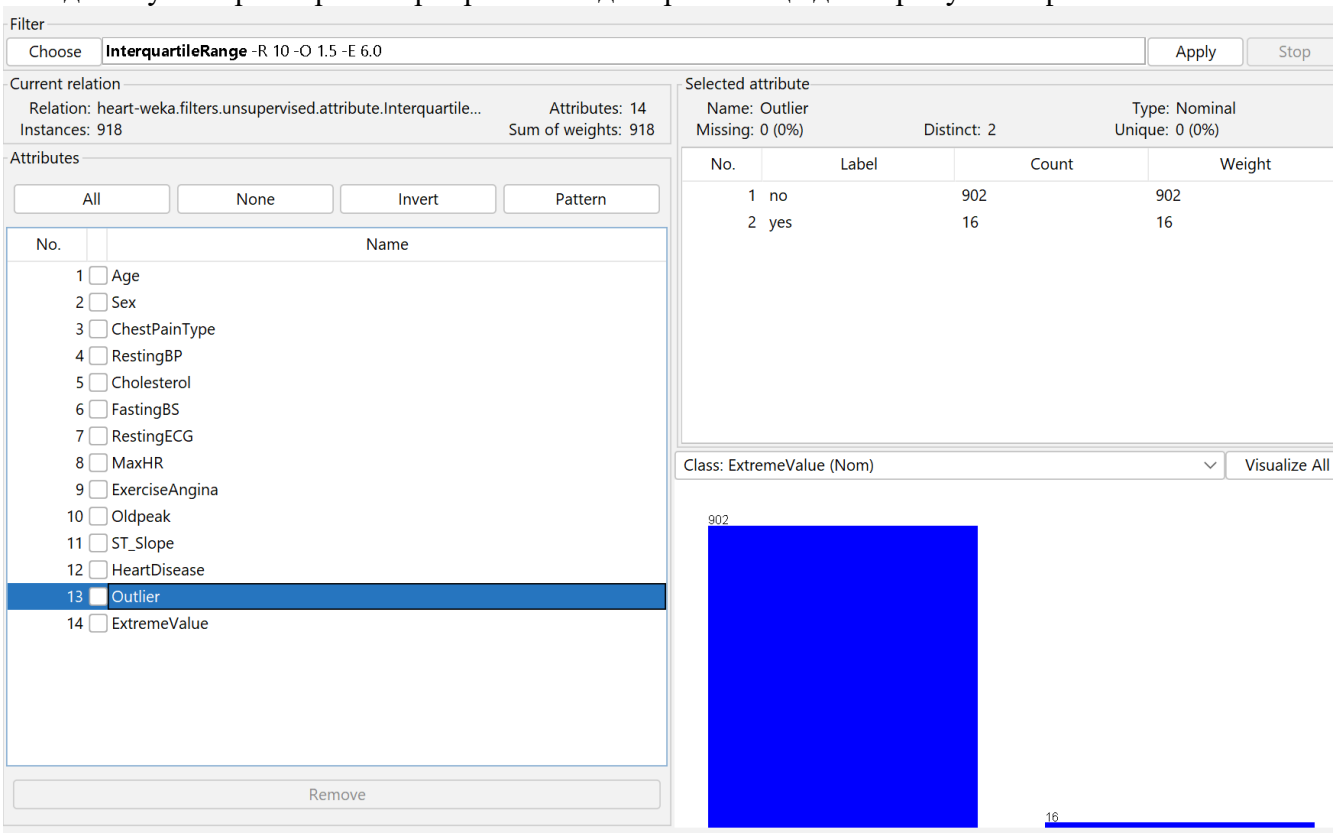
Selected attribute

Name: Cholesterol Type: Numeric
Missing: 0 (0%) Distinct: 222 Unique: 66 (7%)

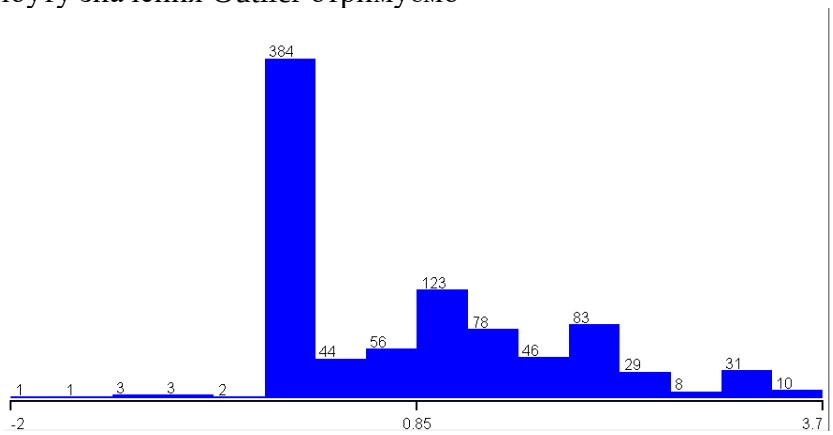
Statistic	Value
Minimum	0
Maximum	603
Mean	198.8
StdDev	109.384

b. У датасеті відсутні пропущені значення, проте присутні викиди у атрибутах RestingBP, Cholesteron, MaxHR і Oldpeak. Відсутні непотрібні або дубльовані атрибути. Проте присутні тесктові значення, які потрібно перетворити на числові.

c. У даному наборі потрібно прибрати викиди. Зробимо це для атрибута Oldpeak.

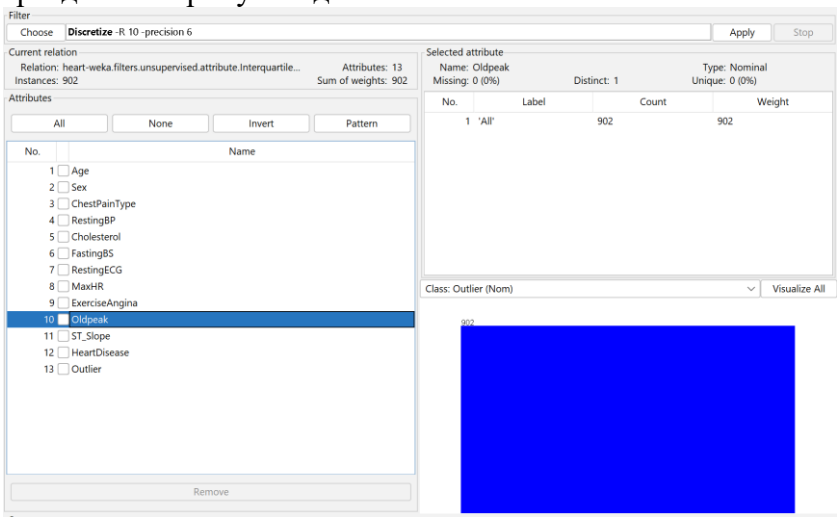


Видаливши із атрибуту значення Outlier отримуємо



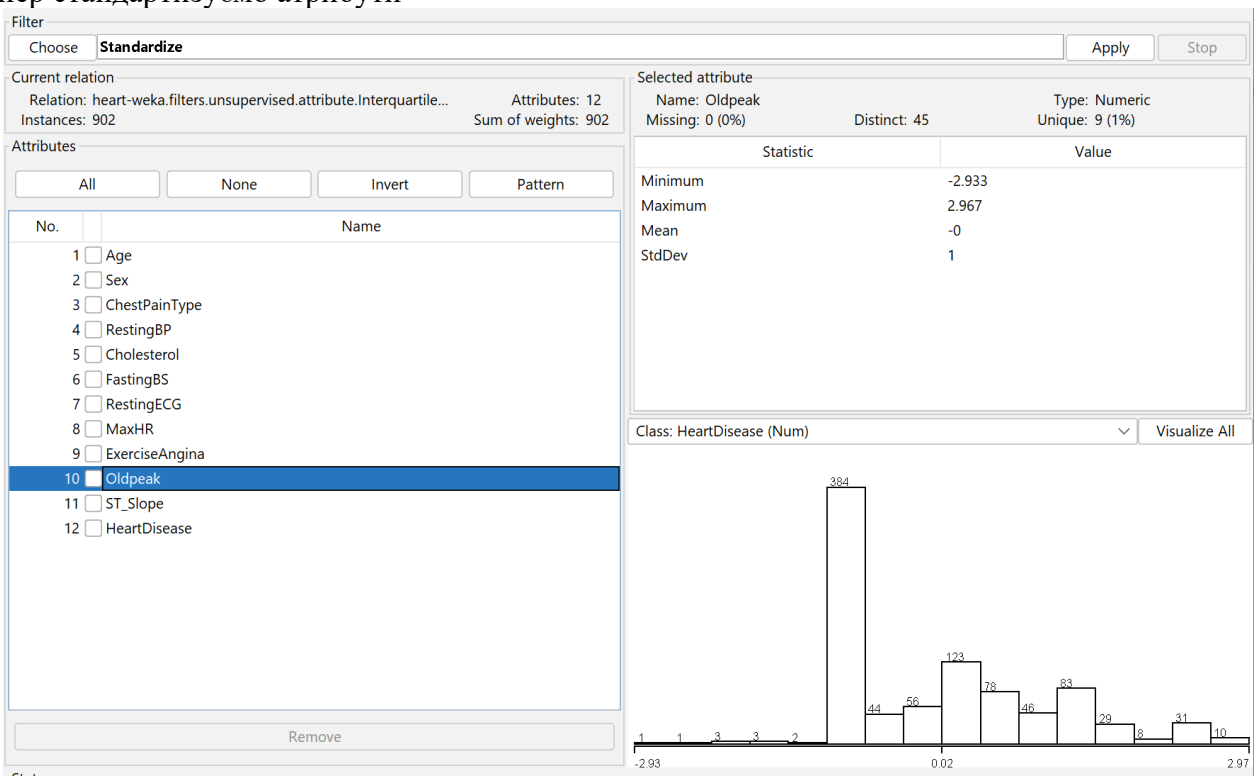
Видалено викиди у атрибуті Oldpeak.

Тепер попробуємо розділити атрибут на декілька сегментів



Як бачимо, у нас не вийшло розділити його на декілька класів, тому не вийде провести зглажування сегментів і тому зразу проведемо загальну стандартизацію.

Тепер стандартизуємо атрибути



Аналіз результатів:

Атрибут Oldpeak містив 16 екземплярів із викидами із 918, що є 2%. Ця фільтрація є незначним покращення датасету, проте якщо провести фільтрацію решти атрибутів із викидами, то результат буде відчутним.

Оскільки в подальшому буде будуватись регресійна модель для виявлення залежностей для хворих не серце, то було вибрано стандартизацію атрибутів.

Висновок:

Ознайомлено студентів з системою WEKA, яка є потужним інструментом для обробки і аналізу даних. Навчився використовувати основні функції цієї системи, зокрема, завантажувати, обробляти і візуалізувати набори даних.

Практичним завданням для вирішення є знайти взаємозв’язок між медичними показниками пацієнта, щоб передбачити, чи можлива у нього серцева хвороба.

Було виявлено викиди у атрибутах RestingBP, Cholesteron, MaxHR і Oldpeak. Провелась фільтрація від викидів атрибуту Oldpeak та стандартизація атрибутів.