

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет "Львівська політехніка"



Ознайомлення з WEKA. Підготовка даних.

МЕТОДИЧНІ ВКАЗІВКИ
до лабораторної роботи № 4

з курсу "Системи інтелектуального аналізу та візуалізації даних"
для студентів за освітньою програмою Комп'ютерні науки (Проектування і
програмування інтелектуальних систем та пристроїв)

Затверджено на засіданні кафедри
"Системи автоматизованого проектування"

Протокол N 1 від 28.08.2023р.

ЛЬВІВ 2023

1. МЕТА РОБОТИ

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.

2. КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

2.1. Вступ

Підготовка даних - це критичний етап у процесі аналізу даних та машинного навчання, оскільки від якості вихідних даних залежить якість навчання моделей та отриманих результатів.

2.1.1. Опис підготовки даних та її значущості у машинному навчанні та аналізі даних:

У Weka ви можете реалізувати ряд дій по підготовці даних, таких як:

1. Завантаження даних: Завантажте набір даних у відповідному форматі файлу (наприклад, ARFF, CSV) і переконайтеся, що структура даних відповідає вимогам.

2. Очищення даних: Ідентифікуєте та виправляєте помилки, аномалії та неправильні дані. Це може включати виправлення синтаксичних помилок та помилкових значень атрибутів.

3. Обробка пропущених значень: Замініть відсутні або неіснуючі значення у даних на середнє значення, моду або медіану (для числових атрибутів) або значення, яке найчастіше зустрічається (для номінальних атрибутів).

4. Видалення або об'єднання атрибутів: Спробуйте зменшити кількість атрибутів, видаливши ті, що містять мало інформації або сильно корелюють, або об'єднайте декілька атрибутів в один.

5. Трансформація атрибутів: Виконайте математичні чи статистичні трансформації числових атрибутів (наприклад, масштабування, нормалізація, логарифмічне перетворення) для підвищення їх значимості та покращення роботи моделі.

6. Векторизація тексту: Якщо набір даних містить текст, перетворіть його на числові або номінальні атрибути, використовуючи методи векторизації тексту.

7. Дискретизація числових атрибутів: При потребі перетворіть числові атрибути на категоріальні або дискретні значення, поділивши їх на декілька дискретних інтервалів або категорій.

8. Розділення набору даних на навчальний та тестовий: Розділіть набір даних на дві частини - одну для навчання моделі, іншу - для перевірки та оцінки її роботи.

Підготовка даних може варіюватись в залежності від проекту. Тому перед застосуванням Weka необхідно детально проаналізувати свій набір даних та визначити, які етапи підготовки найефективніші для конкретного випадку..

2.1.2. Короткий огляд Weka, її переваг та набору інструментів для обробки даних:

Weka (Waikato Environment for Knowledge Analysis) - це платформа відкритого коду для машинного навчання та аналізу даних, що була розроблена на факультеті комп'ютерних систем університету Вайкато в Новій Зеландії. Weka пропонує широкий спектр інструментів і алгоритмів, включаючи основні методи машинного навчання, аналізу даних та статистичного моделювання.

Переваги Weka включають:

- Відкритий код: Weka безкоштовна до використання, а програмісти можуть модифікувати її код для власних потреб.
- Застосування алгоритмів: Weka пропонує велику кількість алгоритмів для машинного навчання та аналізу даних, включаючи класифікацію, регресію, кластеризацію і візуалізацію результатів.
- Робоче середовище: Weka має графічний інтерфейс, що надає зручні візуальні інструменти для аналізу даних, налаштування алгоритмів і візуалізацію результатів.
- Обробка даних: Weka пропонує потужні інструменти для очищення, нормалізації, трансформації та вибору даних, що дозволяє проводити підготовку даних без необхідності використання додаткового програмного забезпечення.
- Weka також пропонує різне програмне забезпечення командного рядка, графічні інструменти та бібліотеки Java для обробки даних, застосовуючи різні фільтри, які допоможуть виконати такі завдання, як:
 - Обробка пропущених значень
 - Стандартизація та нормалізація числових значень
 - Кодування та векторизація категоріальних даних
 - Агрегація та узагальнення даних

2.2. Огляд Weka GUI

2.2.1. Опис робочого середовища Weka: розбір головного меню та доступних опцій.

Weka оснащена графічним інтерфейсом користувача (GUI), який забезпечує доступ до більшості її функцій та інструментів. Головний інтерфейс Weka містить головні елементи меню та ряд налаштувань:



Рис. 1 Початкова сторінка системи Weka.

- "Explorer": це інтерактивний інтерфейс для аналізу даних та застосування алгоритмів машинного навчання. За допомогою Explorer можна управляти наборами даних, обробляти дані за допомогою фільтрів, застосовувати класифікатори, кластери та асоціативні правила, а також створювати візуальні зображення.
- "Experimenter": це інструмент для створення та проведення експериментів з метою порівняння різних алгоритмів на обраних наборах даних. За допомогою Experimenter можна виконувати статистичний аналіз та порівнювати алгоритми за допомогою різних метрик.
- "KnowledgeFlow": це візуальна рамка для проектування та розробки процесів обробки даних та машинного навчання. У KnowledgeFlow користувачі можуть створити візуальне представлення робочих процесів, "переміщуючи" компоненти, які можуть містити вихідні дані, перетворення, моделі та відображення.
- "Workbench": це інтерфейс, що об'єднує Explorer, Experimenter та KnowledgeFlow для більш зручного користування..

2.2.2. Введення в Weka Explorer, Data Preprocessing.

Weka Explorer - це один з найкорисніших та найбільш простих у використанні інструментів Weka, який надає доступ до більшої частини її функціональності, включаючи підготовку даних. Ось основні вкладки в Weka Explorer:

1. "Preprocess": ця вкладка дозволяє завантажувати та обробляти дані, використовуючи ряд фільтрів для наборів даних Weka. Можна налаштовувати параметри фільтру, проводити очищення даних, здійснювати заміну відсутніх значень та трансформувати параметри.
2. "Classify": тут користувач може застосовувати до даних техніки класифікації, отримати оцінку точності моделі та порівняти результати різних алгоритмів.
3. "Cluster": ця вкладка дозволяє працювати з даними за допомогою методів кластеризації, включаючи ієрархічну кластеризацію та метод k-середніх.
4. "Associate": з цієї вкладки користувач може досліджувати асоціативні правила, виявляючи відносини та закономірності між атрибутами в даних.
5. "Select attributes": на цій вкладці можна відібрати найважливіші атрибути за допомогою різних методів відбору, метою якого є зменшення розмірності даних та видалення шумових атрибутів.

6. "Visualize": ця вкладка надає можливість візуалізувати взаємозв'язок між атрибутами у вигляді діаграм розсіювання або графіків, а також аналізувати розподіл значень атрибутів.

Коли вкладки активовані, натискання на них дозволяє переходити між різними екранами, на яких можна виконувати відповідні дії. Незалежно від обраної вкладки, нижня частина вікна залишається видимою.

2.3. Завантаження даних

2.3.1. Формати даних, сумісних з Weka: ARFF, CSV та інші

Weka підтримує кілька форматів даних, що призначені для зручності імпорту та експорту користувачами. Тут наведено деякі з них:

- ARFF (Attribute-Relation File Format): це текстовий формат, специфічний для Weka, який використовується для представлення структурованих даних. Файли формату ARFF складаються з двох основних розділів: заголовок, який описує назви, типи та пов'язані атрибути даних, та розділ даних, у якому зазначені значення записів.

```
% 1. Title: Database for fitting contact lenses
%
% 2. Sources:
%   (a) Cendrowska, J. "PRISM: An algorithm for inducing modular rules",
%       International Journal of Man-Machine Studies, 1987, 27, 349-370
%   (b) Donor: Benoit Julien (Julien@ce.cmu.edu)
%   (c) Date: 1 August 1990
%
@relation contact-lenses

@attribute age {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip {myope, hypermetrope}
@attribute astigmatism {no, yes}
@attribute tear-prod-rate {reduced, normal}
@attribute contact-lenses {soft, hard, none}

@data
%
% 24 instances
%
young,myope,no,reduced,none
young,myope,no,normal,soft
young,myope,yes,reduced,none
young,myope,yes,normal,hard
young,hypermetrope,no,reduced,none
young,hypermetrope,no,normal,soft
young,hypermetrope,yes,reduced,none
```

Рис.2 Структура файлу ARFF.

- CSV (Comma-Separated Values) та TSV (Tab-Separated Values): це текстові формати, представлені в табличному вигляді, відрізняються розділювачами даних. Вони особливо корисні для обміну інформацією між різними застосунками аналізу даних. Як правило, Weka може без проблем працювати з форматами CSV та TSV, але інколи можуть знадобитися невеликі налаштування.

- Деякі інші формати, що підтримуються Weka, включають Excel (XLS, XLSX), JSON, XML та SQL (за допомогою розширень або додаткових модулів).

2.3.2. Основи імпорту даних та перетворення між форматами

Імпорт даних та конвертування між форматами можна виконати у Weka Explorer, з допомогою або без додаткових модулів. Основні кроки:

1. Запустіть Weka та перейдіть до вкладки "Preprocess" в Weka Explorer.
2. Щоб завантажити дані, натисніть кнопку "Open file" та оберіть файл формату, який вам потрібен. Якщо файл у форматі ARFF або CSV, Weka автоматично відкриває його. У разі інших форматів може знадобитись налаштування модулів або розширень.
3. Після завантаження набору даних, ви можете його переглядати й обробляти у Weka Explorer. Якщо потрібно конвертувати дані у інший формат, натисніть кнопку "Save" та виберіть бажаний формат зі списку. Weka створить копію набору даних у новому форматі, яку ви можете використовувати пізніше у інших програмах або Weka-сумісних інструментах.

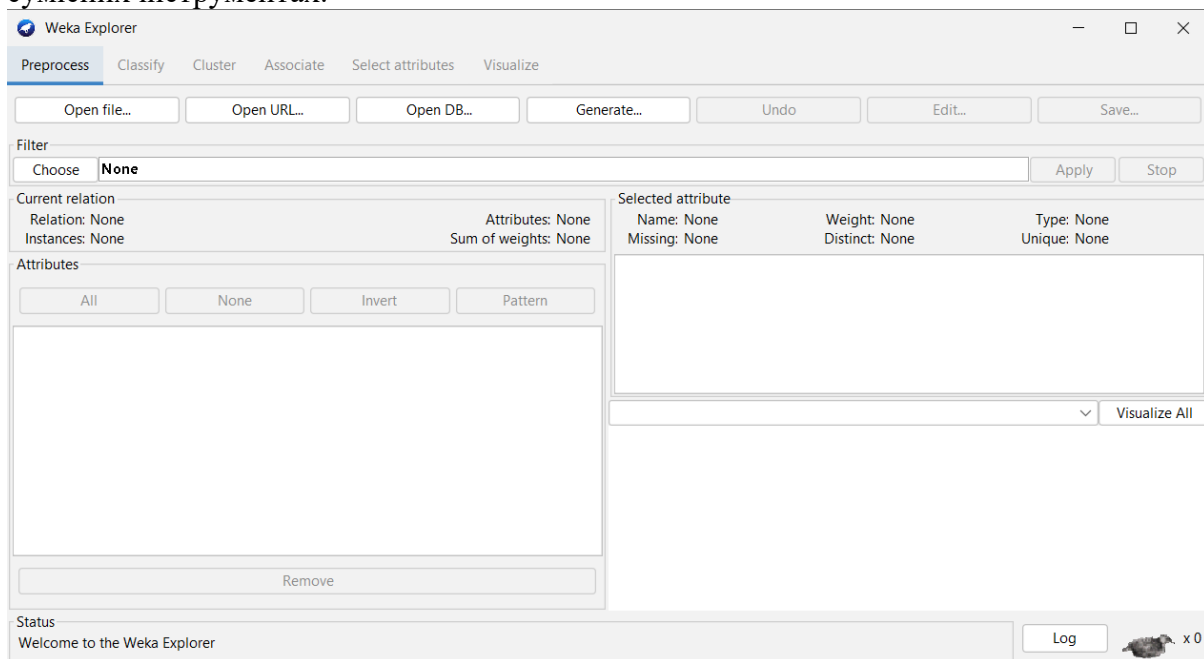


Рис.3 Головна сторінка підсистеми Explorer.

Можна зауважити, що Weka автоматично розпізнає деякі з форматів (наприклад, CSV) під час завантаження. Однак в разі необробленого формату файл імпортувати не вдасться, і можуть виникнути проблеми з кодуванням або заголовками. У такому випадку слід скористатись утилітою "Data Editor" Weka або зконвертувати файл у формат ARFF з відповідними налаштуваннями.

2.4. Очищення даних у Weka

Очищення даних є важливою частиною підготовки даних, оскільки якість вихідних даних впливає на результати машинного навчання. У Weka є ряд фільтрів для очищення даних та виправлення різних проблем.

2.4.1. Використання фільтрів

Weka надає фільтри для перетворення вашого набору даних. Найкращий спосіб побачити, які фільтри підтримуються, і пограти з ними на вашому наборі даних — це використовувати Weka Explorer.

Панель «Filter» дозволяє вибрати фільтр.

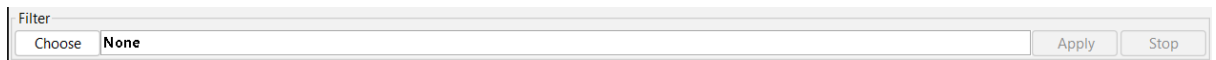


Рис.4 Панель фільтрів Weka для вибору фільтрів даних.

Фільтри діляться на два види:

- Контрольовані (Supervised) фільтри: їх можна застосовувати, але потребують певного контролю користувача. Наприклад, перебалансування екземплярів для класу.
- Неконтрольовані (Unsupervised) фільтри: які можна застосовувати ненаправлено. Наприклад, змініть масштаб усіх значень до діапазону від 0 до 1.

У цих двох групах фільтри далі поділяються на фільтри для атрибутів і екземплярів:

- Фільтри атрибутів: застосовуйте операцію над атрибутами або одним атрибутом за раз.
- Фільтри екземплярів: застосування операції до екземпляра або одного екземпляра за раз.

Після вибору фільтра його назва з'явиться в полі поруч із кнопкою «Choose».

Ви можете налаштувати фільтр, вибравши його назву, після чого відкриється вікно налаштування. Ви можете змінити параметри фільтра і навіть зберегти або завантажити конфігурацію самого фільтра. Це чудово для відтворюваності.

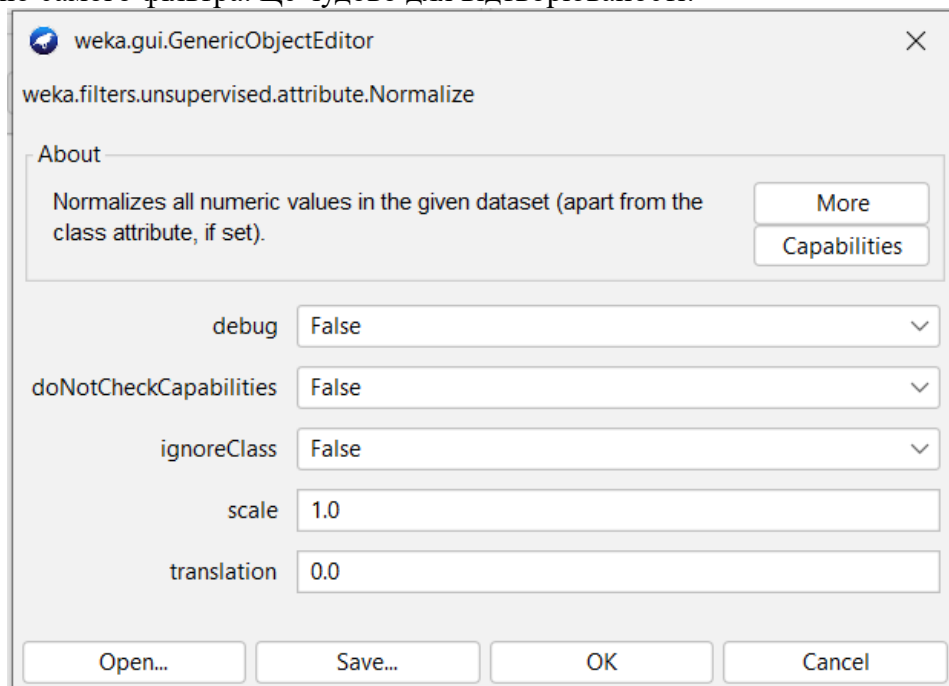


Рис. 5 Конфігурація фільтра даних Weka.

Ви можете дізнатися більше про кожен параметр конфігурації, навівши на нього курсор і прочитавши підказку.

Ви також можете прочитати всі подробиці про фільтр, включаючи конфігурацію, документи та книги для подальшого читання та більше інформації про роботу фільтра, натиснувши кнопку «More».

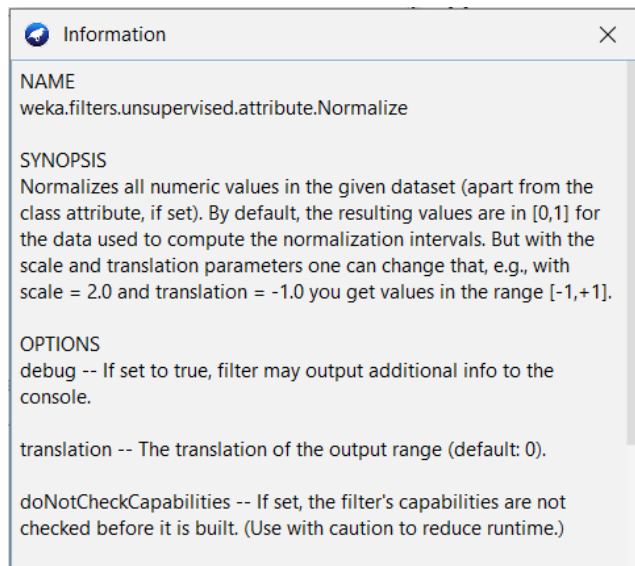


Рис. 6 Фільтр даних Weka. Більше інформації.

Ви можете закрити довідку та застосувати конфігурацію, натиснувши кнопку «OK». Ви можете застосувати фільтр до завантаженого набору даних, натиснувши кнопку «Apply» біля назви фільтра.

2.4.2. виправлення помилок у даних

Фільтр `weka.filters.unsupervised.attribute.ReplaceWithMissingValue` використовується для заміни значень у вибраному атрибуті на пропущені значення ('missing value') на основі встановлених критеріїв. Цей фільтр корисний, коли потрібно розглядати деякі значення атрибутів як відсутні, наприклад, коли значення не є вірними згідно з доменними вимогами або коли потрібно вилучити спотворені дані з подальшого аналізу.

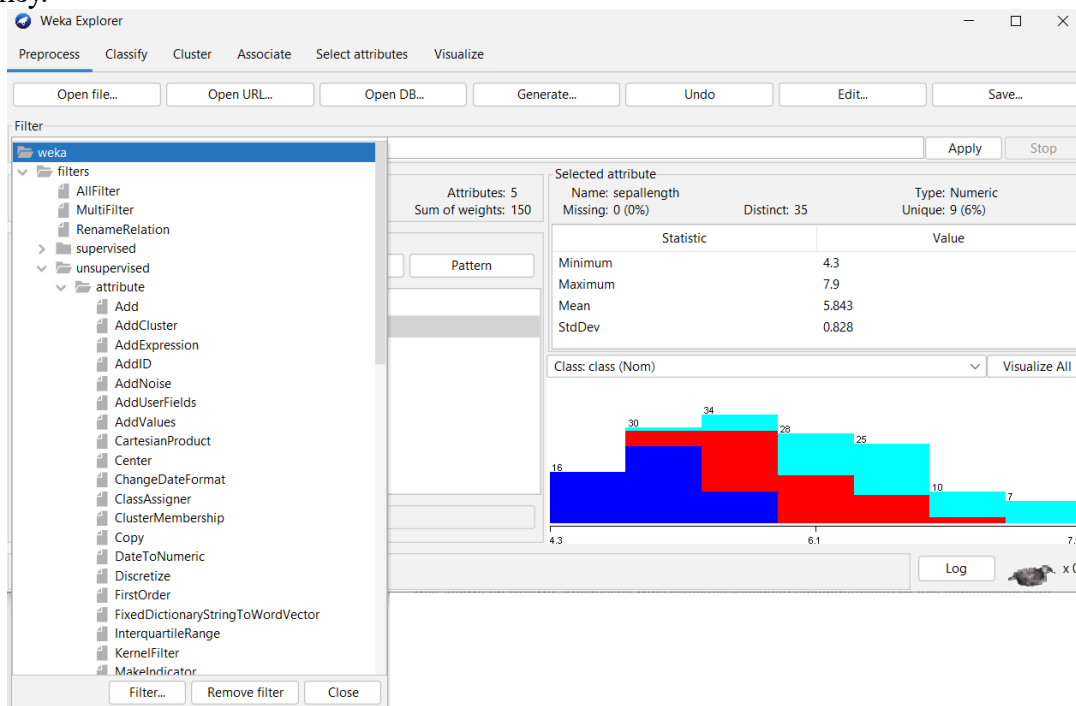


Рис. 7

Для застосування цього фільтра, слід виконати наступні кроки:

1. У вкладці "Preprocess" додайте фільтр `weka.filters.unsupervised.attribute.ReplaceWithMissingValue` (знайдіть його в розділі "filters", підрозділ "unsupervised", а потім "attribute").

2. Після додавання фільтра виберіть його в списку фільтрів та відкрийте його параметри, аби налаштувати фільтр згідно з вашими потребами: виберіть атрибут, у якому ви бажаєте відбуватися заміна значень; вкажіть порогове значення; виберіть, чи буде заміна проводитися для значень, які перевищують поріг, менші за поріг або які дорівнюють порогу.

3. Натисніть "Apply" для застосування фільтра на вашому наборі даних.

Застосуванням цього фільтра можна виправити пропущені або некоректні дані перед використанням методів машинного навчання або проведенням аналізу даних. Проте слід звернути увагу на те, що фільтр не замінює пропущені значення на якісь альтернативні значення; замість цього встановлюється значення "missing value", що може потребувати подальших кроків під час підготовки, таких як використання фільтра "ReplaceMissingValues".

2.4.3. Видалення аномалій

Виявлення та/або видалення аномалій може потребувати використання великої кількості статистичних технік, таких як, наприклад, відхилення, засновані на кластеризації тощо. Weka не має готового фільтра для видалення аномалій, але можна виконати основний аналіз та видалити аномалії вручну або використати сторонні бібліотеки Weka.

Деякі фільтри та опції очищення даних у Weka:

1. Фільтр `"weka.filters.unsupervised.attribute.Remove"`:

Видаляє задані атрибути з набору даних.

Виберіть атрибут (-R) індекси атрибутів, які слід видалити.

2. Фільтр `"weka.filters.unsupervised.attribute.ReplaceMissingValues"`:

Замінює пропущені значення числових атрибутів на середнє значення цього атрибуту та пропущені значення номінальних атрибутів на моду (найчастіше значення) цього атрибуту.

3. Фільтр `weka.filters.unsupervised.attribute.RemoveByName` використовується для видалення атрибутів з набору даних на основі регулярних виразів, які збігаються з назвами атрибутів. Цей фільтр корисний, коли вам потрібно видалити кілька атрибутів, назви яких мають спільний шаблон, або коли потрібно приховати окремі атрибути перед аналізом даних або використанням алгоритмів машинного навчання.

Ці фільтри - лише декілька прикладів можливостей Weka для обробки та очищення даних. Інші фільтри можуть виконувати ряд різних операцій з підготовки даних, включаючи трансформацію атрибутів, кодування категорій та роботу з текстовими даними. Вивчення різних фільтрів та їх параметрів допоможе вам ефективно налаштувати процес підготовки даних для вашого набору даних.

2.5. Обробка пропущених значень у Weka

Обробка пропущених значень є важливою частиною підготовки даних, оскільки моделі машинного навчання зазвичай не можуть добре працювати з відсутніми даними.

При роботі з даними, в яких є пропущені значення атрибутів для деяких екземплярів, існують наступні стратегії поведінки.

1. Відкинути екземпляри з пропущеними значеннями. Такий підхід застосовується насамперед для даних, у яких відсутнє значення цільового атрибута (для задач класифікації).

2. Заповнити пропущені значення вручну.

3. Застосувати глобальну константу (наприклад, «Unknown»).

4. Використати деяке статистично розраховане по всій вибірці значення (середнє арифметичне, медіану, моду).

5. Використати статистичне значення, розраховане для примірників, що відносяться до того ж класу, як і розглянутий екземпляр.

6. Використати найбільш ймовірне значення для атрибута. Це значення може бути розраховане за допомогою регресії, дерева рішень або інших математичних підходів.

У Weka є декілька методів та фільтрів для обробки пропущених значень.

Фільтр `"weka.filters.unsupervised.attribute.ReplaceMissingValues"` можна використовувати для автоматичної заміни пропущених значень обидвох числових та номінальних атрибутів. Для числових атрибутів пропущені значення замінюються середнім значенням, а для номінальних значень - модою (найбільш поширеним значенням). Важливо зауважити, що Weka не має вбудованої можливості замінити пропущені значення медіаною, але таку функцію можна реалізувати власноруч або використовувати сторонні пакети для отримання медіани.

2.6. Трансформація та відбір атрибутів у Weka

2.6.1. Методи стандартизації (масштабування), нормалізації та дискретизації числових атрибутів:

Масштабування: відомо, що різні атрибути можуть мати різні діапазони значень, що може вплинути на результати машинного навчання. Тому масштабування даних до спільного діапазону - це важливий крок препроцесингу. Фільтр `"weka.filters.unsupervised.attribute.Standardize"` може стандартизувати всі числові атрибути до середнього значення 0 та стандартного відхилення 1.

Стандартизація припускає, що ваші дані мають гаусівський (дзвоноподібний) розподіл. Це не обов'язково має бути правдою, але ця методика ефективніша, якщо ваш розподіл атрибутів є гаусовим.

Ви можете гаусівським всі атрибути у своєму наборі даних за допомогою Weka, вибравши фільтр `Standardize` і застосувавши його до свого набору даних.

Ви можете скористатися наведеним нижче набором кроків, щоб стандартизувати свій набір даних:

1. Відкрийте Weka Explorer

2. Завантажте набір даних.

3. Натисніть кнопку «Choose», щоб вибрати фільтр і виберіть `unsupervised.attribute.Standardize`.

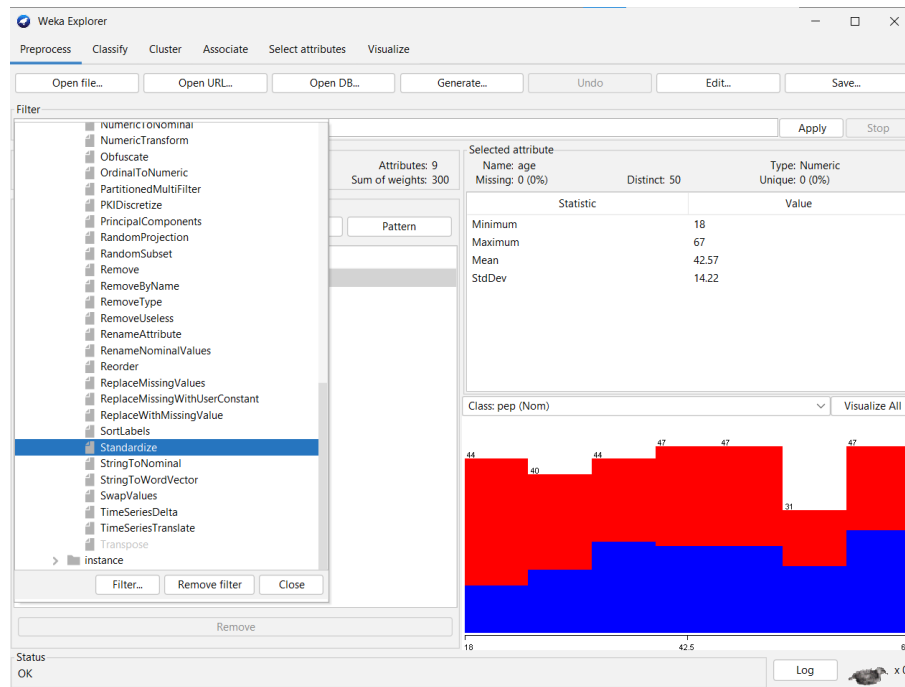


Рис.8 Вибір фільтра стандартизації даних.

4. Натисніть кнопку «Apply», щоб нормалізувати свій набір даних.

5. Натисніть кнопку «Save» та введіть назву файлу, щоб зберегти стандартизовану копію вашого набору даних.

Перегляд деталей кожного атрибута у вікні «Вибраний атрибут» дасть вам впевненість, що фільтр був успішним і кожен атрибут має середнє значення 0 і стандартне відхилення 1.

Selected attribute		
Name: age	Distinct: 50	Type: Numeric
Missing: 0 (0%)		Unique: 0 (0%)
Statistic	Value	
Minimum	-1.728	
Maximum	1.718	
Mean	0	
StdDev	1	

Рис.9 Стандартизований розподіл даних Weka.

Стандартизація корисна, коли ваші дані мають різні масштаби, а алгоритм, який ви використовуєте, робить припущення про те, що дані мають розподіл Гауса, як-от лінійна регресія, логістична регресія та лінійний дискримінантний аналіз.

Нормалізація — хороший метод, який можна використовувати, коли ви не знаєте розподіл своїх даних або якщо знаєте, що розподіл не є гаусівським.

Ви можете нормалізувати всі атрибути у своєму наборі даних за допомогою Weka, вибравши фільтр Нормалізація та застосувавши його до свого набору даних.

Ви можете використати такий рецепт, щоб нормалізувати свій набір даних:

1. Відкрийте Weka Explorer.
2. Завантажте набір даних.

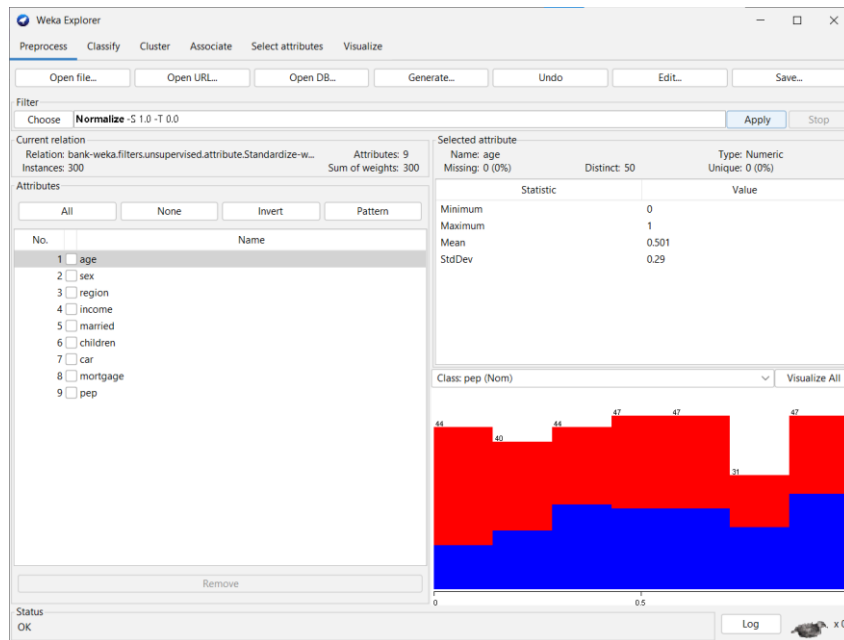


Рис.10 Завантажений набір даних про банки у Weka Explorer.

3. Натисніть кнопку «Choose», щоб вибрати фільтр і виберіть `unsupervised.attribute.Normalize`.

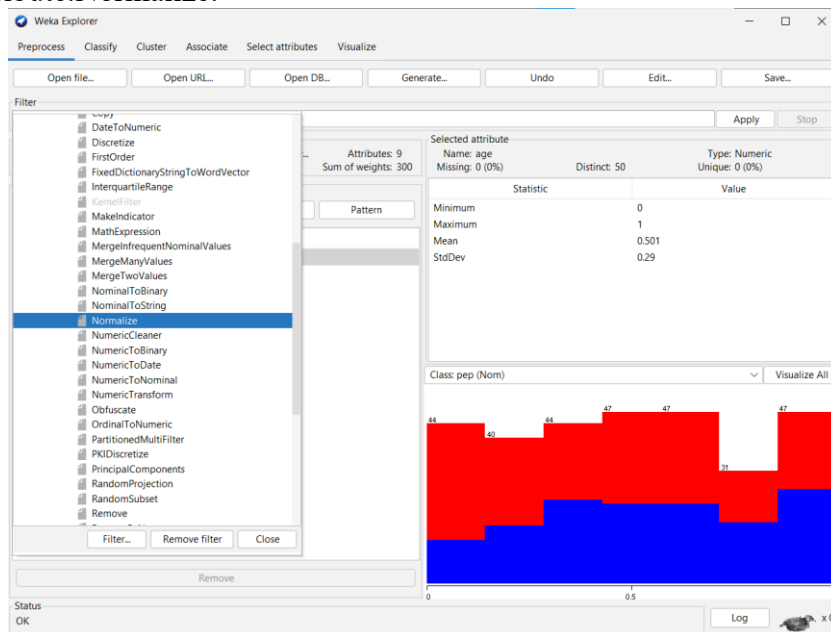


Рис.11 Weka фільтр вибору нормалізації даних.

4. Натисніть кнопку «Apply», щоб нормалізувати свій набір даних.

5. Натисніть кнопку «Save» та введіть назву файлу, щоб зберегти нормалізовану копію вашого набору даних.

Перегляд деталей кожного атрибута у вікні «Вибраний атрибут» дасть вам впевненість, що фільтр був успішним і кожен атрибут було змінено в діапазоні від 0 до 1.

Selected attribute		
Name: age		Type: Numeric
Missing: 0 (0%)	Distinct: 50	Unique: 0 (0%)
Statistic		Value
Minimum		0
Maximum		1
Mean		0.501
StdDev		0.29

Рис. 12 Нормалізований розподіл даних у Weka.

Ви можете використовувати інші масштаби, наприклад від -1 до 1, що корисно при використанні опорних векторних машин.

Нормалізація корисна, коли ваші дані мають різні масштаби, а алгоритм, який ви використовуєте, не робить припущень щодо розподілу ваших даних, і який заснований на обчисленні відстаней (наприклад k-найближчих сусідів) і штучних нейронних мереж.

Дискретизація числових атрибутів є обов'язковою і необхідною у разі застосування алгоритмів інтелектуального аналізу, що працюють тільки з категоріальними атрибутами. Крім того, алгоритми, що працюють з числовими атрибутами часто дають кращі результати або ж працюють швидше, якщо значення атрибутів попередньо приведені до дискретної форми.

Методи дискретизації можуть бути класифіковані за двома параметрами:

- чи використовується в них інформація про класи: дискретизація з учителем (supervised discretization) або дискретизація без вчителя (unsupervised discretization);
- в якому напрямку відбувається дискретизація:
 - зверху-вниз (дискретизація починається з однієї або декількох точок поділу, а далі отримані інтервали рекурсивно розбиваються; метод розбиття);
 - знизу-вгору (спочатку всі значення атрибуту розглядаються як потенційні точки поділу, а далі сусідні значення рекурсивно об'єднуються, утворюючи інтервали; об'єднання).

2.6.2. Векторизація тексту для перетворення текстових атрибутів:

Одним із відомих методів векторизації тексту є "мішок слів" (Bag-of-Words). Це може бути реалізовано за допомогою фільтра "weka.filters.unsupervised.attribute.StringToWordVector" у Weka. Цей фільтр перетворює текстові атрибути на числові, де кожен атрибут відповідає наявності або частоті слова в тексті. Фільтр може бути налаштований для контролю токенизації, стоп-слів, стемінга та інших параметрів.

2.6.3. Опис фільтрів та стратегій відбору атрибутів для зменшення розміру набору даних та видалення шуму та неінформативних атрибутів:

Фільтр "weka.filters.supervised.attribute.AttributeSelection": цей фільтр дозволяє автоматично відбирати атрибути на основі різних методів оцінки атрибутів (наприклад,

GainRatio, Correlation) та алгоритмів пошуку (наприклад, BestFirst, GreedyStepwise, Ranker). Атрибути, що мають найбільше значення для класифікації, обираються та зберігаються, а менш значущі атрибути видаляються.

Фільтр "weka.filters.unsupervised.attribute.RemoveUseless": цей фільтр видаляє усі атрибути, які через відсутність варіативності або інші критерії вважаються некорисними для класифікації. Фільтр може бути налаштований для задання максимальної дисперсії або інших параметрів.

2.7. Розділення даних на тренувальні та тестові набори у Weka

2.7.1. Методи секвенційного та переміщеного розділення наборів даних:

Секвенційне розділення: дані діляться на послідовні частини, наприклад, 70% даних призначається для тренування, а 30% - для тестування. Цей метод можна використовувати, коли відсутні явне впорядкування даних або наявність часових залежностей між ними.

Переміщене (Stratified) розділення: Даний метод має на меті гарантувати, що розподіл міток класів буде збережено в обох тренувальних та тестових наборах, особливо якщо класи є незбалансованими. У випадку переміщеного розділення, спочатку дані розподіляються за класами, а потім кожен клас ділиться у відповідному співвідношенні між тренувальним та тестовим наборами.

2.7.2. Опис налаштування параметрів для розділення даних та розв'язання проблем разом з Weka:

Розділення набору даних на тренувальний та тестовий набори у Weka можна здійснити двома основними способами. Один з них полягає у використанні модуля Experimenter, а інший - за допомогою коду Java (за допомогою Weka API).

Використання Weka Experimenter:

1. Запустіть Weka та оберіть "Experimenter" з головного меню.

2. Натисніть кнопку "New" зліва вгорі, щоб створити новий експеримент.

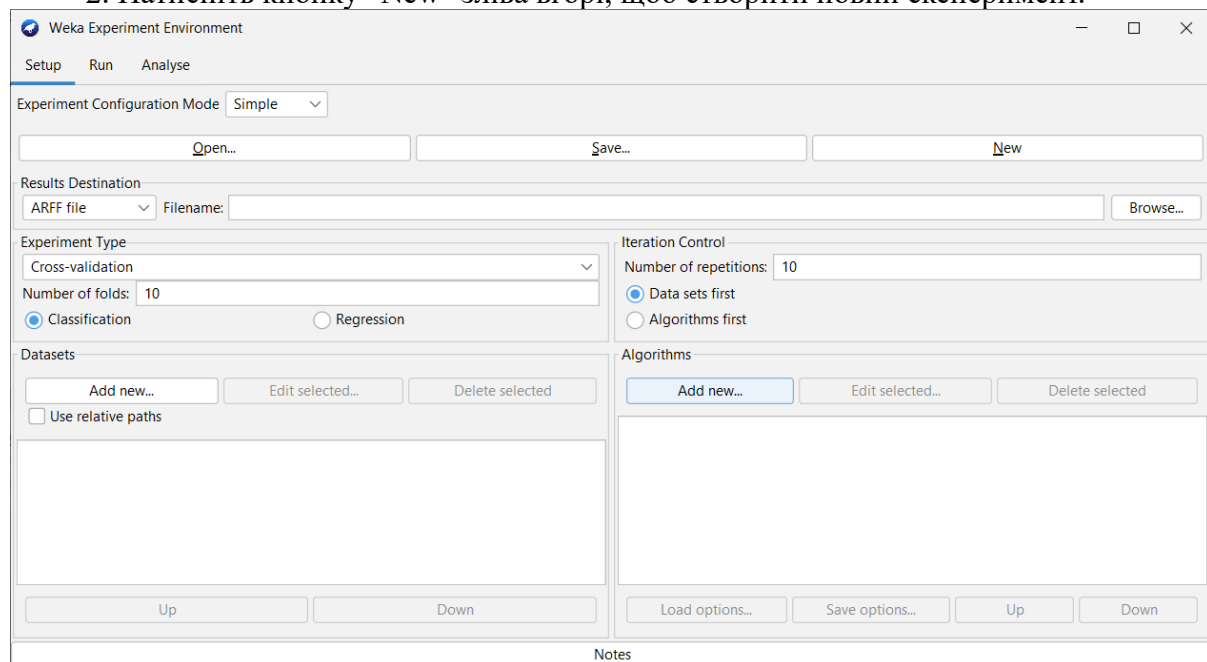


Рис. 13

3. У розділі "Datasets" натисніть кнопку "Add new..." та оберіть файл з даними, який потрібно розділити.
4. У розділі "Algorithms" додайте свою модель або алгоритм.
5. Для визначення розміру-тестового набору, встановіть "Percentage Split" у полі "Test mode" на потрібне значення (наприклад, 70 % для тренування).
6. Збережіть налаштування та виконайте експеримент.
7. В результаті ви отримаєте співвідношення точності моделі на основі обраної установки розділення.

2.8. Практичний приклад попередньої обробки даних

Проект аналізу даних зосереджений на класифікації клієнтів банківської установи на основі історичних даних про кредитних позиках. Мета проекту - визначити, чи є позичальник потенційно надійним або ненадійним. Для цього ми будемо використовувати Weka для попередньої обробки даних, зокрема завантаження, очищення, обробки пропущених значень, трансформації атрибутів та розділення даних.

Покрокові вказівки щодо роботи з Weka для завантаження, очищення, обробки пропущених значень, трансформації, розділення даних та інших кроків попередньої обробки даних:

1. Завантаження даних: Запустіть Weka Explorer і відкрийте вкладку "Preprocess". Виберіть "Open file" та завантажте свій файл із даними.

2. Очищення даних: Перевірте, чи містять дані викиди або нетипові значення. Наприклад, якщо значення кредитного стану має бути у межах від 300 до 850, але ви помічаєте значення 10000, є зміст замінити або видалити аномальні значення. Для цього можна використати фільтри типу RemoveRange або видалити поле вручну у Data Grid.

3. Оброблення пропущених значень: Виберіть фільтр ReplaceMissingValues зі списку фільтрів у вкладці "Preprocess" та додайте його до обраних фільтрів. Виберіть фільтр в обраному списку та натисніть "Apply" для заповнення пропущених значень в даних за допомогою середнього або медіани, залежно від типу атрибуту.

4. Трансформація ознак: Відповідно до потреб аналізу даних, вам може знадобитися кодування категоріальних даних або нормалізація числових атрибутів. Для кодування категоріальних даних використовуйте фільтри типу NominalToBinary або StringToNominal. Для нормалізації числових атрибутів використовуйте фільтри Normalize або Standardize.

5. Розділення даних: Для ефективного проведення машинного навчання вам слід розділити дані на дві частини: навчальний набір (наприклад, 70% даних) та тестовий набір (наприклад, 30% даних). Використайте фільтр "removePercentage" для цього, або вручну збережіть деякі проценти записів у окремих файлах.

Після завершення попередньої обробки даних ви помітите, що дані стали готовими для наступного етапу - проведення класифікації та інших аналітичних методів машинного навчання. Тепер ви можете використати Weka Explorer для побудови класифікаційних моделей, кластеризації, вибору ознак та оцінки результатів на перетворених і попередньо оброблених даних.

Цей приклад ілюструє деякі базові операції з попередньої обробки даних, які можуть бути виконані за допомогою WEKA. Набір даних, що використовується у цьому прикладі, - це "банківські дані".

Дані містять наступні поля:

id	a unique identification number
age	age of customer in years (numeric)

sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO)
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

Після завантаження даних WEKA розпізнає атрибути та під час сканування даних обчислить базову статистику для кожного атрибута. На лівій панелі на рис. 14 показано список розпізнаних атрибутів, тоді як на верхніх панелях вказано назви базового відношення (або таблиці) і поточного робочого відношення (які спочатку однакові).

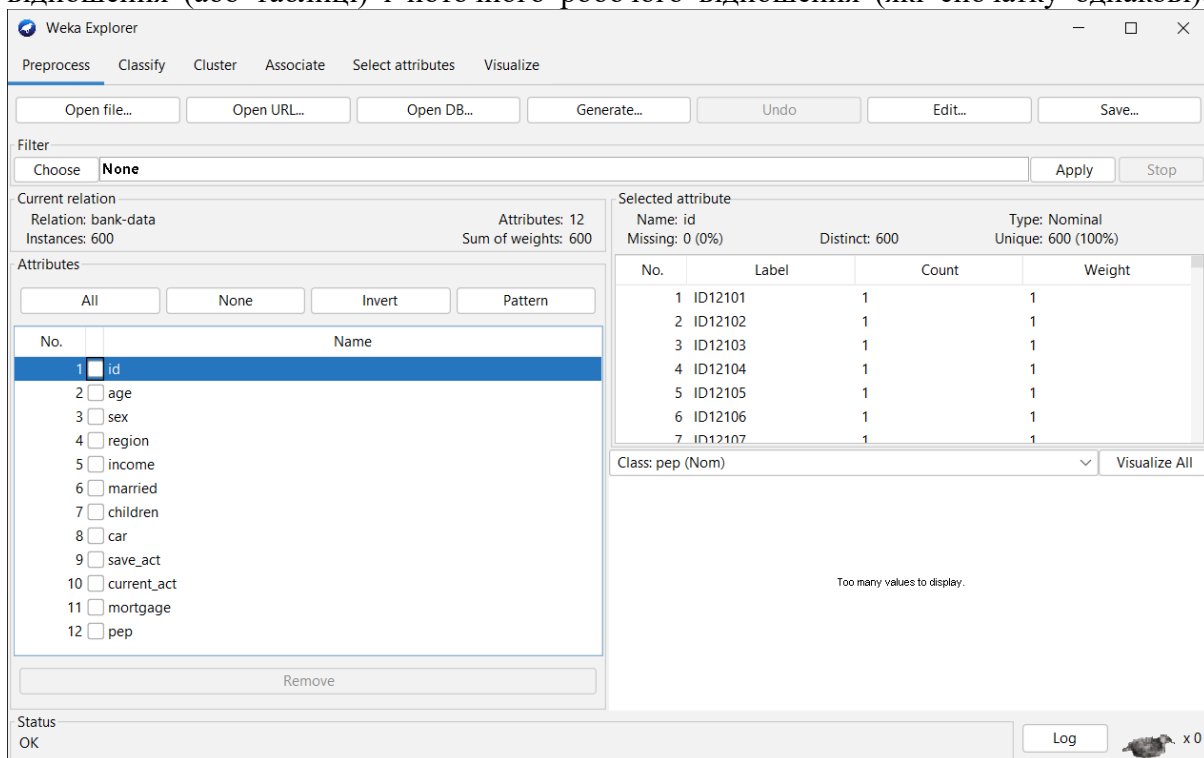


Рис. 14

Якщо натиснути на будь-який атрибут на панелі ліворуч, відобразиться основна статистика щодо цього атрибута. Для категоріальних атрибутів показано частоту для кожного значення атрибута, тоді як для безперервних атрибутів ми можемо отримати мінімальне, максимальне, середнє значення, стандартне відхилення тощо. Як приклад дивіться рис. 15 і 16 нижче, на яких показано результати вибору «age» і атрибуту «children» відповідно.

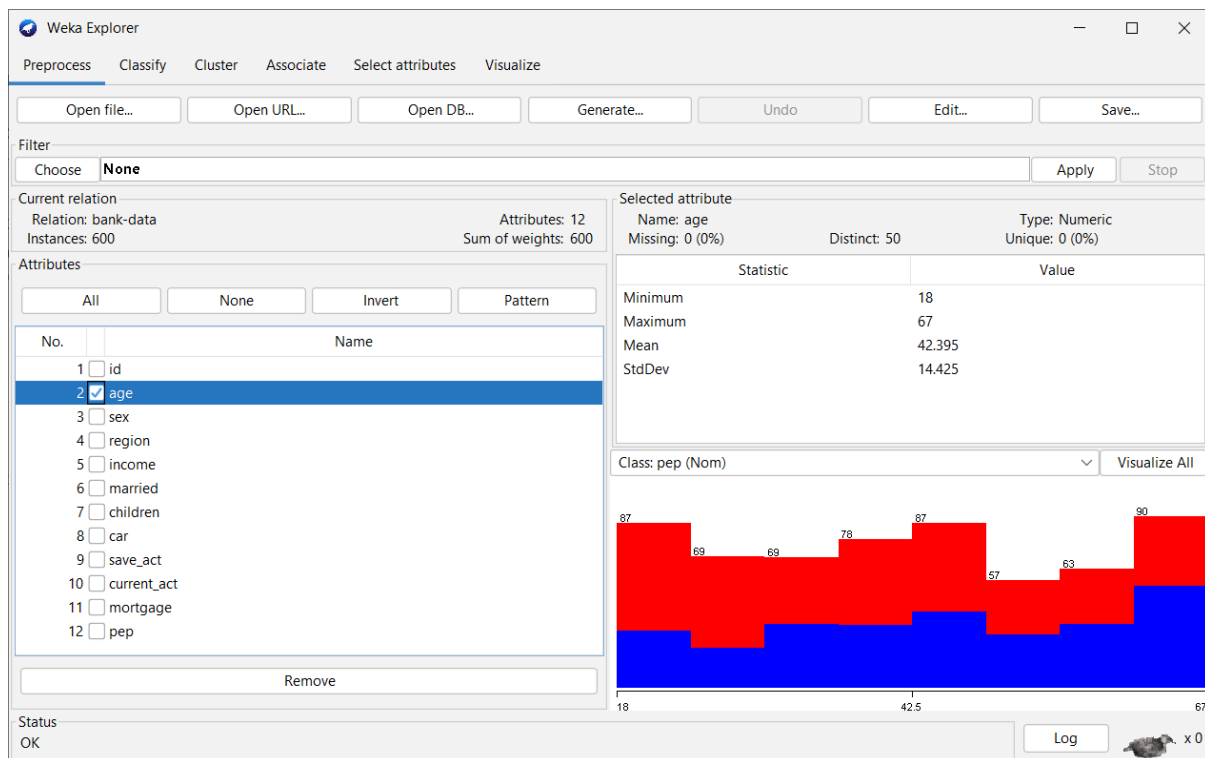


Рис. 15

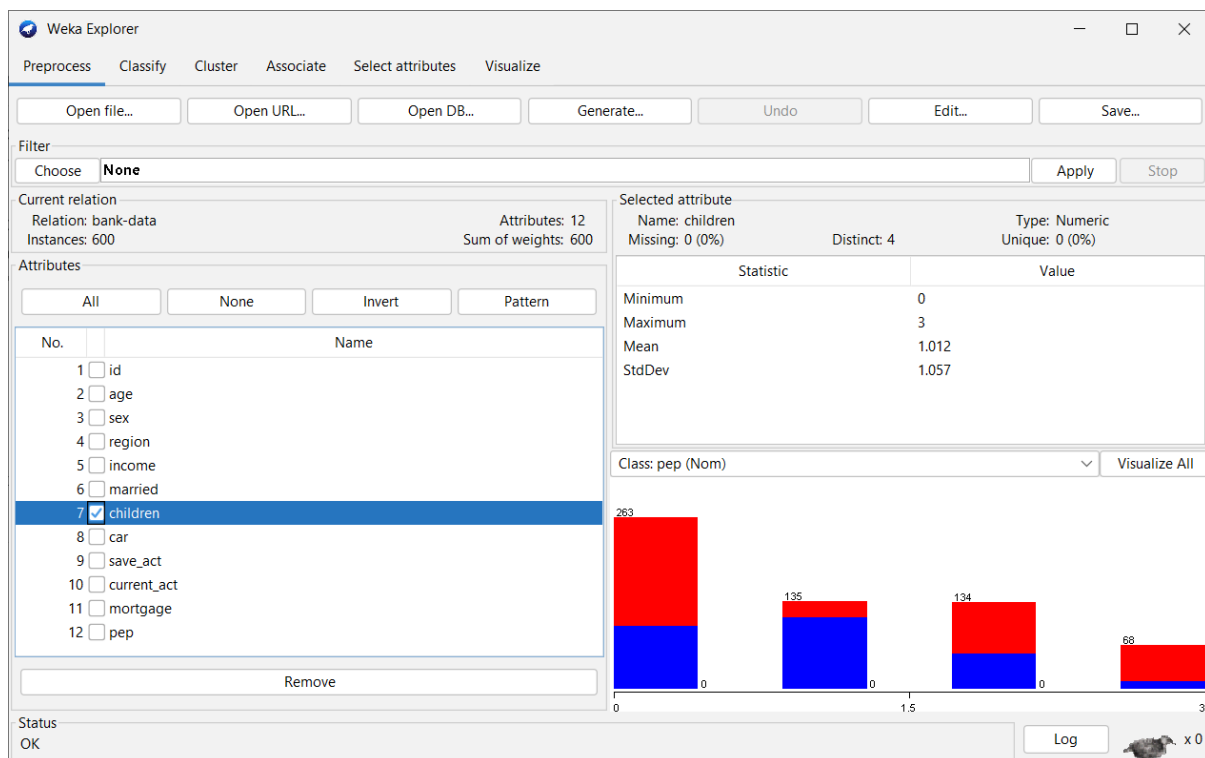


Рис. 16

Зауважте, що візуалізація на правій нижній панелі є формою перехресної таблиці між двома атрибутами. Наприклад, на рис. 16 панель візуалізації за замовчуванням перехресно табулює «children» з атрибутом «pep» (за замовчуванням другим атрибутом

є останній стовпець файлу даних). Ви можете вибрати інший атрибут за допомогою спадного списку.

У нашому прикладі кожен запис унікально ідентифікується ідентифікатором клієнта (атрибут "id"). Нам потрібно видалити цей атрибут перед кроком аналізу даних. Ми можемо зробити це за допомогою фільтрів атрибутів. Виберіть фільтр «weka.filters.unsupervised.attribute.Remove, як показано на рис. 17.

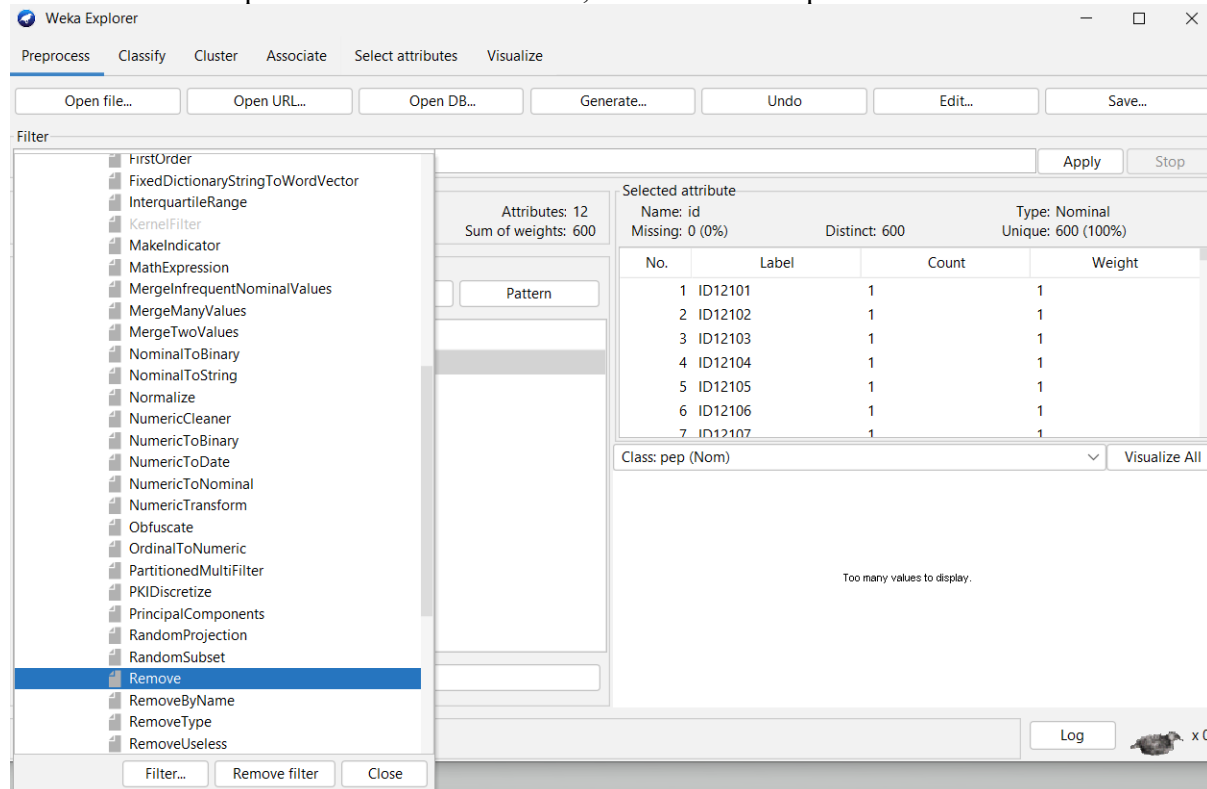


Рис. 17

Далі клацніть текстове поле безпосередньо праворуч від кнопки «Вибрати». У діалоговому вікні, що з'явиться, введіть індекс атрибута, який потрібно відфільтрувати (це може бути діапазон або список, розділений комами). У цьому випадку ми вводимо 1, який є індексом атрибута «id» (див. ліву панель). Переконайтеся, що для параметра "invertSelection" встановлено значення false (інакше все, крім атрибута 1, буде відфільтровано). Потім клацніть «ОК». Тепер у вікні фільтра ви побачите «Видалити -R 1» (див. рис.18).

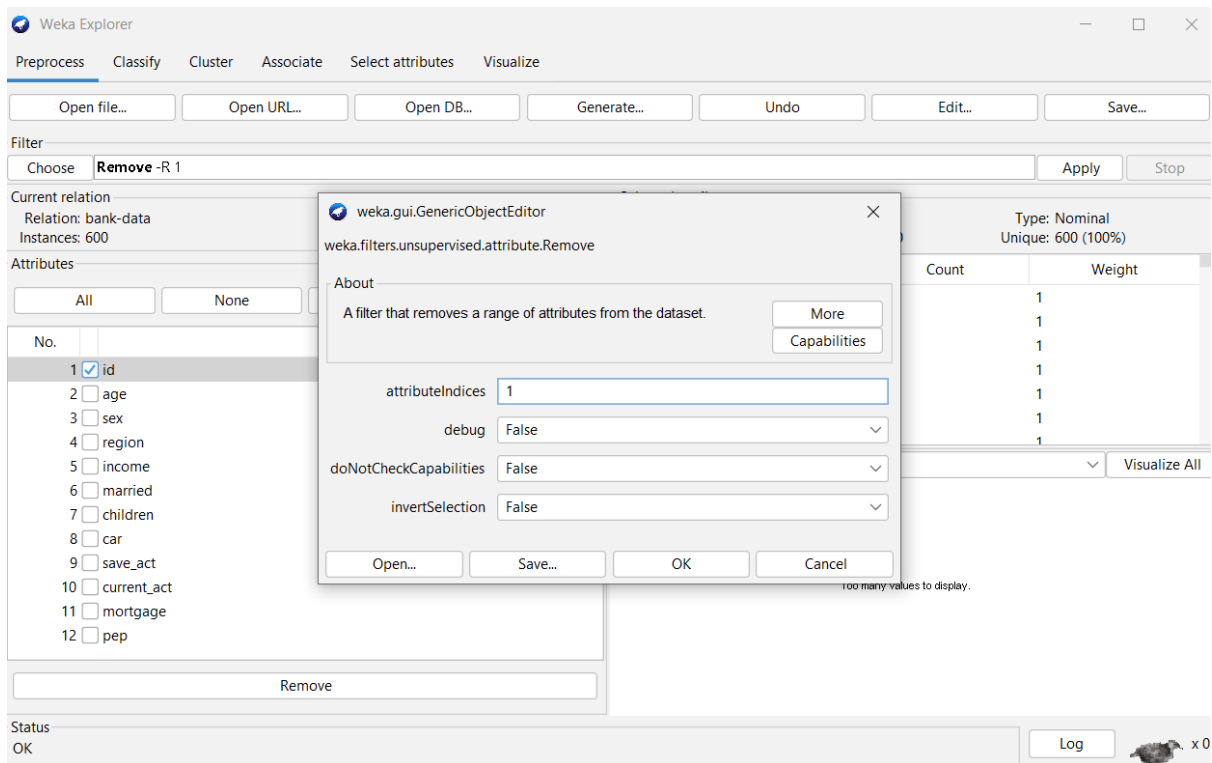


Рис. 18

Натисніть кнопку «Apply», щоб застосувати цей фільтр до даних. Це видалить атрибут «id» і створить новий робочий зв'язок (ім'я якого тепер містить деталі застосованого фільтра). Результат зображено на рис.19.

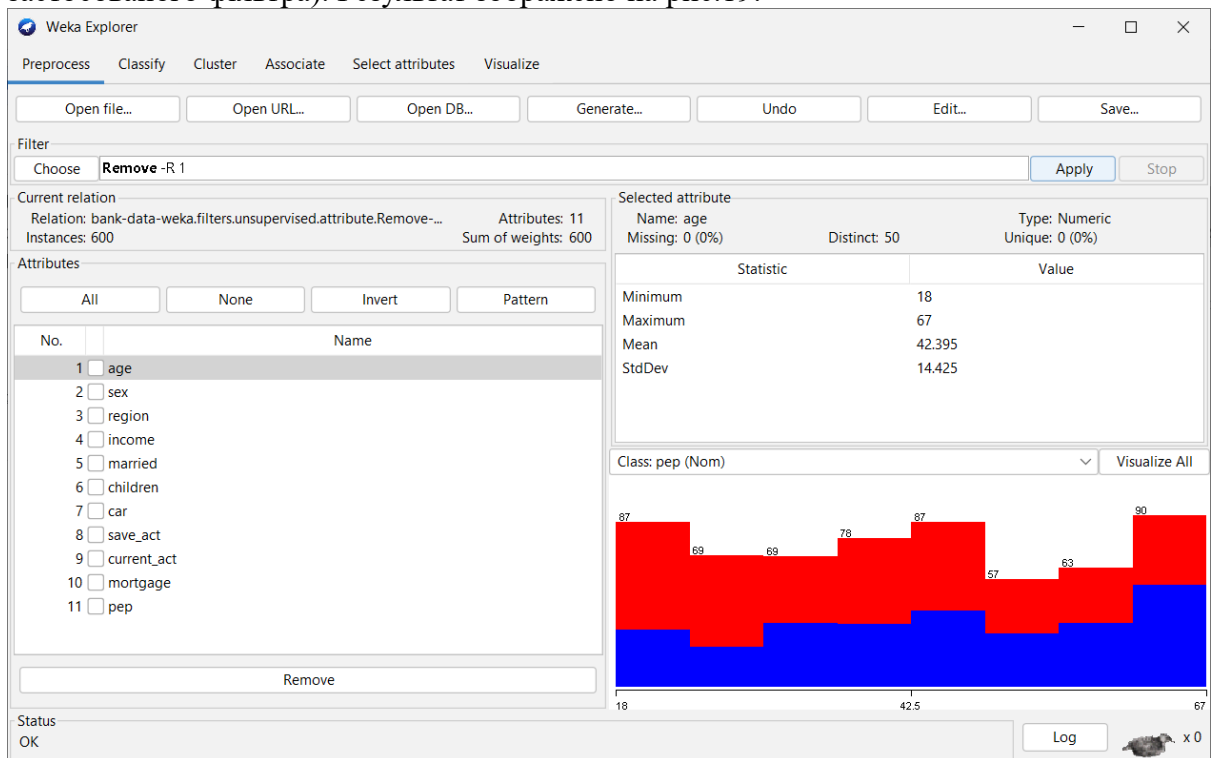


Рис. 19

Тепер можна застосувати додаткові фільтри до нових робочих відносин. Однак у цьому прикладі ми збережемо наші проміжні результати як окремі файли даних і розглядатимемо кожен крок як окремий сеанс WEKA. Щоб зберегти новий робочий зв'язок як файл ARFF, натисніть кнопку «Зберегти» на верхній панелі. Тут, як показано в діалоговому вікні «зберегти» (див. рис.20), ми збережемо новий зв'язок у файлі «bank-data-R1.arff».

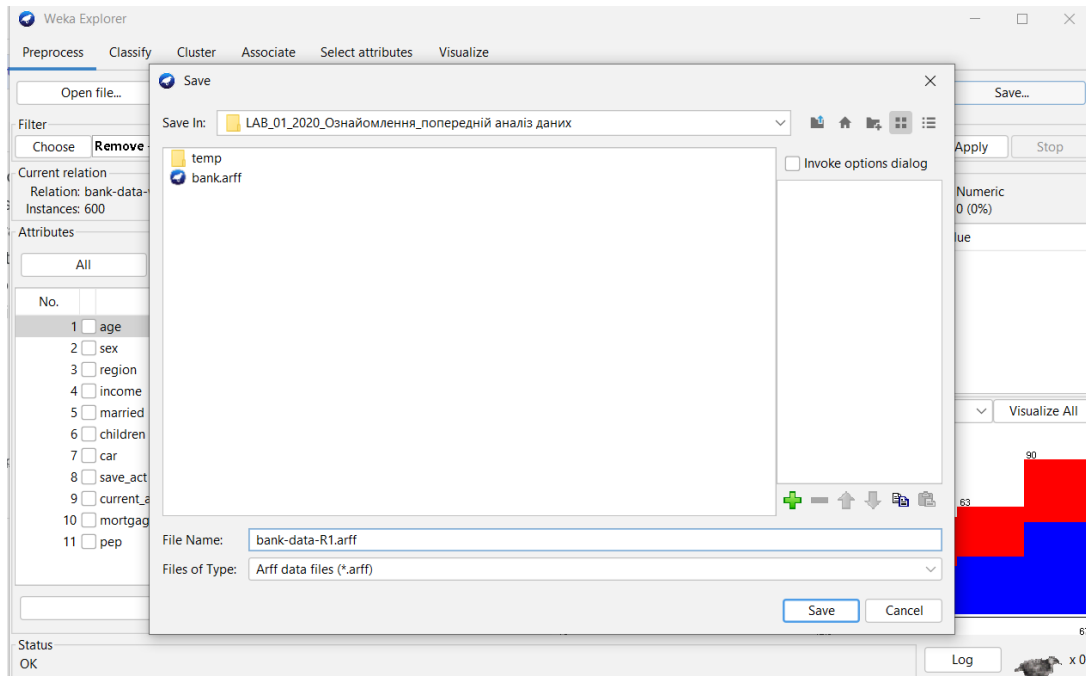


Рис. 20

На рис. 21 показано верхню частину нового створеного файлу ARFF (у TextPad).

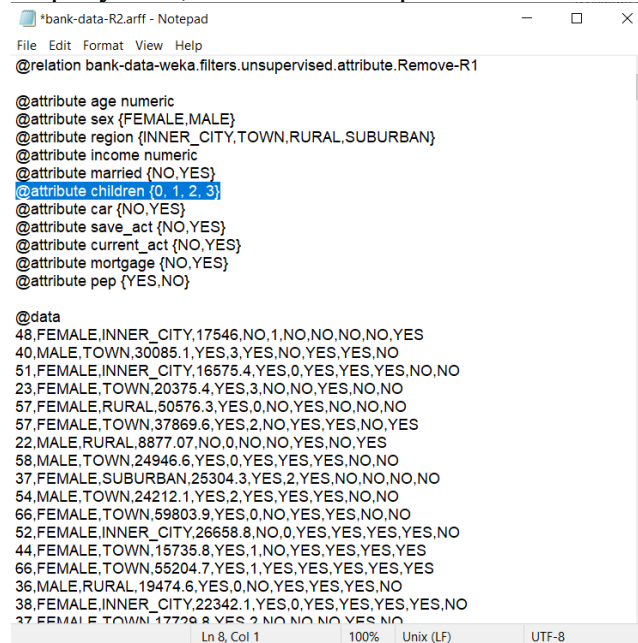
```
bank-data-R1.arff - Notepad
File Edit Format View Help
@relation bank-data-weka.filters.unsupervised.attribute.Remove-R1

@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children numeric
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

@data
48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,NO,YES
40,MALE,TOWN,30085.1,YES,3,YES,NO,YES,YES,NO
51,FEMALE,INNER_CITY,16575.4,YES,0,YES,YES,YES,NO,NO
23,FEMALE,TOWN,20375.4,YES,3,NO,NO,YES,NO,NO
57,FEMALE,RURAL,50576.3,YES,0,NO,YES,NO,NO,NO
57,FEMALE,TOWN,37869.6,YES,2,NO,YES,YES,NO,YES
22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES
58,MALE,TOWN,24946.6,YES,0,YES,YES,YES,NO,NO
37,FEMALE,SUBURBAN,25304.3,YES,2,YES,NO,NO,NO,NO
54,MALE,TOWN,24212.1,YES,2,YES,YES,YES,NO,NO
66,FEMALE,TOWN,59803.9,YES,0,NO,YES,YES,NO,NO
52,FEMALE,INNER_CITY,26658.8,NO,0,YES,YES,YES,YES,NO
44,FEMALE,TOWN,15735.8,YES,1,NO,YES,YES,YES,YES
66,FEMALE,TOWN,55204.7,YES,1,YES,YES,YES,YES,YES
36,MALE,RURAL,19474.6,YES,0,NO,YES,YES,YES,NO
38,FEMALE,INNER_CITY,22342.1,YES,0,YES,YES,YES,YES,NO
37,FEMALE,TOWN,17729.8,YES,2,NO,NO,NO,YES,NO
```

Рис. 21

Зверніть увагу, що в новому наборі даних атрибут "id" і всі відповідні значення в записах видалено. Також зауважте, що Weka автоматично визначила правильні типи та значення, пов'язані з атрибутами, як зазначено в розділі «Attributes» файлу ARFF.



```
*bank-data-R2.arff - Notepad
File Edit Format View Help
@relation bank-data-weka.filters.unsupervised.attribute.Remove-R1

@attribute age numeric
@attribute sex {FEMALE,MALE}
@attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
@attribute income numeric
@attribute married {NO,YES}
@attribute children {0, 1, 2, 3}
@attribute car {NO,YES}
@attribute save_act {NO,YES}
@attribute current_act {NO,YES}
@attribute mortgage {NO,YES}
@attribute pep {YES,NO}

@data
48,FEMALE,INNER_CITY,17546,NO,1,NO,NO,NO,YES
40,MALE,TOWN,30085,1,YES,3,YES,NO,YES,YES,NO
51,FEMALE,INNER_CITY,16575,4,YES,0,YES,YES,YES,NO,NO
23,FEMALE,TOWN,20375,4,YES,3,NO,NO,YES,NO,NO
57,FEMALE,RURAL,50576,3,YES,0,NO,YES,NO,NO,NO
57,FEMALE,TOWN,37869,6,YES,2,NO,YES,YES,NO,YES
22,MALE,RURAL,8877.07,NO,0,NO,NO,YES,NO,YES
58,MALE,TOWN,24946,6,YES,0,YES,YES,YES,NO,NO
37,FEMALE,SUBURBAN,25304,3,YES,2,YES,NO,NO,NO,NO
54,MALE,TOWN,24212,1,YES,2,YES,YES,YES,NO,NO
66,FEMALE,TOWN,59803,9,YES,0,NO,YES,YES,NO,NO
52,FEMALE,INNER_CITY,26658,8,NO,0,YES,YES,YES,YES,NO
44,FEMALE,TOWN,15735,8,YES,1,NO,YES,YES,YES,YES
66,FEMALE,TOWN,55204,7,YES,1,YES,YES,YES,YES,YES
36,MALE,RURAL,19474,6,YES,0,NO,YES,YES,YES,NO
38,FEMALE,INNER_CITY,22342,1,YES,0,YES,YES,YES,YES,NO
37,FEMALE,TOWN,17729,8,YES,2,NO,NO,NO,YES,NO
```

Рис. 22

2.9. Висновки

Підготовка даних є критично важливим етапом у процесі машинного навчання та аналізу даних, оскільки якість вхідних даних безпосередньо впливає на точність та віддачу від моделей та аналітичних інструментів. Weka надає потужні можливості препроцесингу, що дозволяє з легкістю обробляти дані, готуючи їх для наступних етапів аналізу та машинного навчання. Використовуючи Weka, користувачі можуть виконувати різні завдання препроцесингу, такі як очищення даних, видалення пропущених значень, нормалізація та стандартизація числових атрибутів, кодування категоріальних змінних та зменшення розмірності. Успішне застосування препроцесингу даних забезпечує кращі передумови для побудови ефективних моделей машинного навчання та отримання достовірних результатів аналізу даних.

3. ЛАБОРАТОРНЕ ЗАВДАННЯ ТА ВАРІАНТИ ІНДИВІДУАЛЬНИХ ЗАВДАНЬ

1. Визначте та охарактеризуйте набір даних

а. Ви можете вибрати набір даних зі списку загальнодоступних наборів даних у сховищі машинного навчання UCI або в розділі наборів даних на Weka (посилання подані нижче). Ви також можете самостійно досліджувати набори даних з інших джерел.

б. Використовуючи вкладки попередньої обробки даних та візуалізації, проведіть детальний опис вибірки даних. Вкажіть:

- яке практичне завдання вирішується;
- скільки примірників у вибірці;
- атрибути, які характеризують екземпляри вибірки, їхні типи та опис;

- чи є екземпляри з відсутніми значеннями, чи є викиди у даних;
- який атрибут є цільовим, які значення він приймає, скільки екземплярів кожного класу у вибірці.

2. Дослідження та попередня обробка даних

а. Виберіть один атрибут та обговоріть відповідні міри центральної тенденції та дисперсії для атрибуту. Використовуйте підмножину значень атрибутів (на власний вибір) із набору даних і обчисліть середнє, медіану, режим, діапазон, квартилі та дисперсію для атрибута.

б. Обговоріть питання якості даних набору даних. Чи існують (потенційні) проблеми з певними атрибутами даних? Які відповіді на ці питання якості?

с. Обговоріть кілька методів попередньої обробки даних, які, ймовірно, необхідні для набору даних. Наприклад, чи потрібне згладжування даних або зменшення даних, і що було б відповідною технікою. Виберіть один атрибут і використовуйте підмножину значень атрибутів, щоб зробити наступне: 1) розділіть їх на відповідну кількість сегментів з рівною частотою, а також розділенням однакової ширини, 2) використовуйте засоби згладжування за допомогою bin, щоб згладити дані на основі розділення, 3) нормалізуйте атрибут на основі мінімальної нормалізації та нормалізації z-оцінки. Прокоментуйте, який метод ви бажаєте використовувати для розділення, згладжування та нормалізації для даного атрибута.

3. Дослідить можливості Weka

а. Завантажте набір даних у Weka.

б. Дослідить функції візуалізації та попередньої обробки («візуалізувати», «попередню обробку» та «вибрати атрибут»), використовуючи ваш набір даних. Обговоріть нову інформацію, яку ви знайшли при візуалізації даних, випробуваннях технік та отриманих результатів.

с. Зауваження: Weka використовує формат даних, який називається ARFF. Більшість наборів даних у сховищі UCI можна завантажити у розділі наборів даних на веб-сайті Weka у форматі ARFF. Однак для інших наборів даних, що представляють необроблені дані проблемної області, спочатку потрібно перевести їх у формат ARFF.

Бібліотеки наборів даних:

1. UCI Machine Learning Repository: <https://archive.ics.uci.edu/>
2. Datasets section at Weka: <http://www.cs.waikato.ac.nz/ml/weka/>

4. КОНТРОЛЬНІ ЗАПИТАННЯ

1. Що таке інтелектуальний аналіз даних?
2. Для чого використовується програма WEKA, які її можливості?
3. Яке призначення модулів Explorer, Knowledge Flow, Experimenter, Command-Line Interface?
4. Опишіть формат arff файлу.
5. Основні типи змінних, які використовуються у Інтелектуальному аналізі даних?
6. Призначення вкладок в модулі Explorer: Preprocess, Classify, Cluster, Associate, Select Attributes, Visualize.

7. Що таке генеральна сукупність і вибірка? Якими властивостями повинні володіти дані? Що таке репрезентативна вибірка?
8. Що розуміють під фільтрацією у Weka? В чому різниця між фільтрами атрибутів та фільтрами екземплярів? В чому різниця між unsupervised та supervised фільтрами?
9. Що таке якість даних? Яка мета підготовки даних до аналізу? Які завдання входять в підготовку даних?
10. Який атрибут даних називають цільовим?
11. Що таке значимий та незначимий атрибут? Що таке відбір атрибутів?
12. За допомогою яких фільтрів можна виконати наступні завдання підготовки даних:
 - перетворити тип атрибута;
 - нормалізувати значення числового атрибута;
 - знайти та замінити відсутні значення в даних;
 - видалити всі екземпляри даних з заданим значенням атрибута;
 - створити новий атрибут;
 - виконати відбір атрибутів;
 - знайти викиди в даних;
 - створити підвибірку даних.

5. ЗМІСТ ЗВІТУ

1. Тема і мета роботи.
2. Індивідуальне завдання до роботи.
3. Результати виконання завдань.
4. Висновки (відображують результати виконання роботи та їх критичний аналіз).