Міністерство освіти і науки України
Національний університет водного господарства та природокористування
Кафедра комп'ютерних наук та прикладної математики

**04-01-75М**

# МЕТОДИЧНІ ВКАЗІВКИ

до виконання лабораторних робіт з навчальної дисципліни

«Системи інтелектуального аналізу даних»

для здобувачів вищої освіти другого (магістерського) рівня

за освітніми програмами «Бізнес-аналітика»

та «Управління персоналом та економіка праці»

спеціальності *051 Економіка*

денної та заочної форм навчання

Частина 1

Рекомендовано науково-методичною радою з якості ННІЕМ
Протокол № 9 від 15.06.2023 р.

Рівне – 2023

Методичні вказівки до виконання лабораторних робіт з навчальної дисципліни «Системи інтелектуального аналізу даних» для здобувачів вищої освіти другого (магістерського) рівня за освітніми програмами «Бізнес-аналітика» та «Управління персоналом та економіка праці» спеціальності 051 Економіка денної, заочної форм навчання. [Електронне видання] / Прищепа О. В. – Рівне : НУВГП, 2023. – 20 с.

Укладач:
Прищепа О. В. – к.ф.-м.н, доцент кафедри комп'ютерних наук та прикладної математики.

Відповідальний за випуск:
Турбал Ю. В. – д.т.н., професор, завідувач кафедри комп'ютерних наук та прикладної математики.

Керівники груп забезпечення:
Олійник О. О., к.е.н, доцент
Мазур Н. О., к.е.н, доцент

**ЗМІСТ**

# ВСТУП

Системи інтелектуального аналізу даних є важливим курсом у процесі формування сучасного фахівця з економіки. Студенти знайомляться з технологіями Data Mining, її методами, інструментальними засобами та особливостями застосування. Розглянуті у курсі системи доцільно використовувати для розв'язання задач соціально-економічного прогнозування та планування розвитку промислових галузей, підприємств, інших служб, що забезпечують функціонування міст, областей та регіонів.

Метою вивчення дисципліни «Системи інтелектуального аналізу даних» є засвоєння студентами технології Data Mining, призначених для обробки великих обсягів даних та визначення корисних на практиці закономірностей; застосування теоретичних відомостей процесу аналізу даних за допомогою технології Data Mining, вміння оперувати при цьому комбінацією вивчених методів, здійснювати вибір ефективних методів та підходів до аналізу даних.

Основними завданнями, що мають бути вирішені при вивченні дисципліни, є: формування у студентів знань технологій Data Mining; формування навиків застосування методів інтелектуального аналізу даних; вміння вибрати ефективні методи розв'язування задач; вміння аналізувати отримані результати.

## Лабораторна робота №1

**Тема:** Ознайомлення з програмним пакетом WEKA для виконання інтелектуального аналізу даних.

**Теоретичні відомості:**

**WEKA** (Waikato Environment for Knowledge Analysis є вільним програмним забезпеченням для інтелектуального аналізу даних та машинного навчання, що написане на Java в університеті Ваїкато, Нова Зеландія, поширюється по ліцензії **GNU** General Public License. Перша версія **WEKA** вийшла в 1993 році, вона була написана на різних мовах програмування. Лише у 1997 році було прийнято рішення переписати програму на мову Java.

Основним інтерфейсом користувача програмного пакету WEKA є Explorer (рис. 1), що містить графічний користувальницький інтерфейс для роботи з файлами даних спеціального типу і формування результатів у вигляді таблиць та графіків.


Рис. 1. Програмний пакет WEKA.

Зокрема, інтерфейс Explorer (рис. 2) містить наступні панелі: панель попереднього опрацювання, панель класифікації, панель асоціації, панель кластеризації, панель вибору атрибутів, панель візуалізації, що забезпечують виконання наступних завдань аналізу даних: підготовка даних (попередня обробка, відбір ознак), кластеризація, класифікація, зокрема, дерева рішень, пошук асоціативних правил, регресійний аналіз, візуалізація результатів.

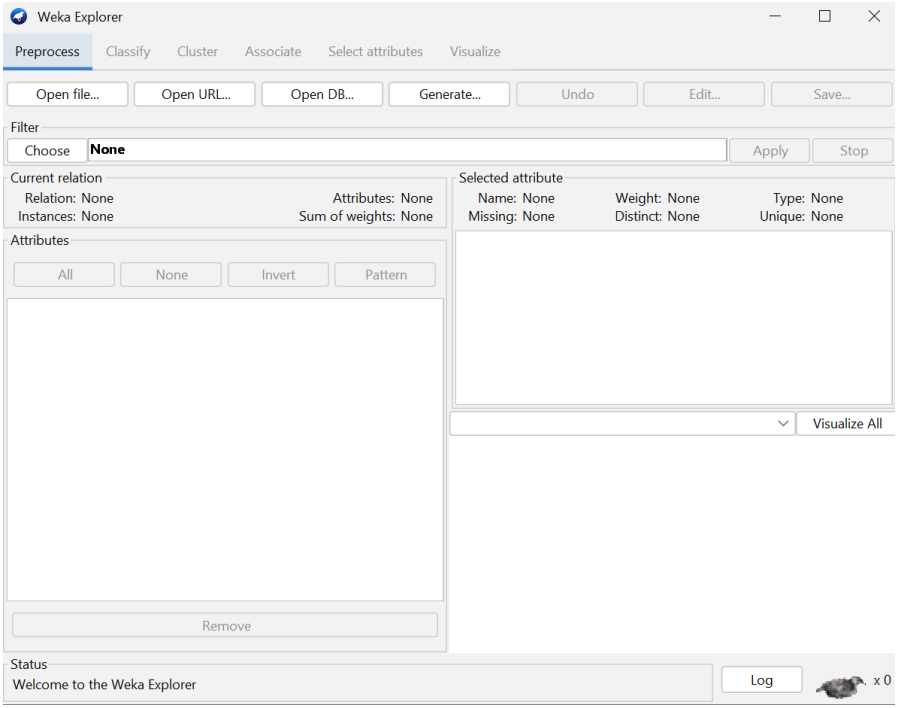


Рис. 2. Панелі інтерфейсу Explorer.

Для завантаження даних у **WEKA**, слід передбачити формат, зрозумілий для цього програмного пакету. Форматом для завантаження даних в **WEKA** є **ARFF** (Attribute-Relation File Format),

який спочатку визначає тип завантажуваних даних, а потім вказує власне дані.

У ARFF-файлі вказується назва і тип даних для кожного стовпця таблиці (атрибуту), а потім дані по рядках для кожного об’єкту, розділені комою.

Для прикладу наведемо ARFF-файл з даними про ціни на будинки (рис. 3):

houseSize – розмір будинку;
numbedrooms – кількість кімнат;
wardrobe – гардероб;
addbathroom – додаткова ванна кімната;
sellingPrice – ціна продажу.

```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE numbedrooms NUMERIC
@ATTRIBUTE wardrobe NUMERIC
@ATTRIBUTE addbathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
352,6,0,0,205000
324,5,1,1,224900
403,5,0,1,197900
239,4,1,0,189900
220,4,0,1,195000
353,6,1,1,325000
298,5,0,1,230000
```

Рис. 3. ARFF-файл даних для завантаження у WEKA.

**Завдання.**

1. Завантажити програмний пакет WEKA за посиланням https://waikato.github.io/weka-wiki/downloading_weka/
2. Дослідити функціональні можливості програмного пакету WEKA.
3. Сформувати дані у форматі ARFF відповідно до обраної теми, що буде досліджуватися.
4. Використати панель попереднього опрацювання для редагування даних.

**Питання для контролю рівня знань:**

1. Що таке інтелектуальний аналіз даних?
2. Які Ви знаєте методи інтелектуального аналізу даних?
3. Який формат даних використовують при роботі з програмним пакетом WEKA?

## Лабораторна робота №2

**Тема:** Інтелектуальний аналіз даних: метод регресійного аналізу з використанням програмного пакету WEKA.

**Теоретичні відомості:**

Регресійний аналіз – найпростіший метод інтелектуального аналізу даних, проте є при цьому найменш ефективним. Побудова моделі регресійного аналізу є дуже зручним прикладом для початку роботи з пакетом **WEKA**, що демонструє можливості інтелектуального аналізу, зокрема, використання великих наборів даних. Одновимірний випадок – це є найпростіша модель аналізу даних, що враховує один незалежний параметр для прогнозування залежної змінної. Але більшість практичних задач спрямовані на побудову моделі регресійного аналізу для прогнозування значення однієї залежної змінної на основі даних відомих значень декількох незалежних параметрів.

Для створення регресійної моделі в **WEKA** потрібно на закладці **Preprocess** відкрити ARFF-файл з даними (використати **Open file**), що буде використовуватися при дослідженні.

1) Скористаємося панеллю класифікації (**Classify**), рис.1. Визначаємо тип моделі для аналізу, для цього натискаємо **Choose** і розгортаємо меню **functions**, обираємо опцію **LinearRegression**. У разі регресійного аналізу додатково потрібна опція **Use training set**. У цьому випадку **WEKA** створить модель на базі даних із завантаженого ARFF-файлу.
2) При створенні моделі важливим етапом є вибір залежної змінної, тобто стовпець, в якому знаходиться невідоме значення, яке потрібно розрахувати. Відразу після секції **Test options** знаходиться список, що розкривається, в якому вам потрібно вибрати залежний параметр. Натискаємо кнопку **Start.**
3) У вікні **Classifier output** отримаєте модель лінійної регресії та значення коефіцієнту кореляції **Correlation coefficient.**

Для прикладу, на основі файлу (рис.3) буде побудована регресійна модель (рис. 4):

$$\begin{aligned} \text{sellingPrice} = &(-346{,}9665 * \text{houseSize}) + \\ &+ (57613{,}4404 * \text{numbedrooms}) + \\ &+ (39259{,}8389 * \text{wardrobe}) + \\ &+ (49397{,}8767 * \text{addbathroom}) - 7718{,}517. \end{aligned}$$

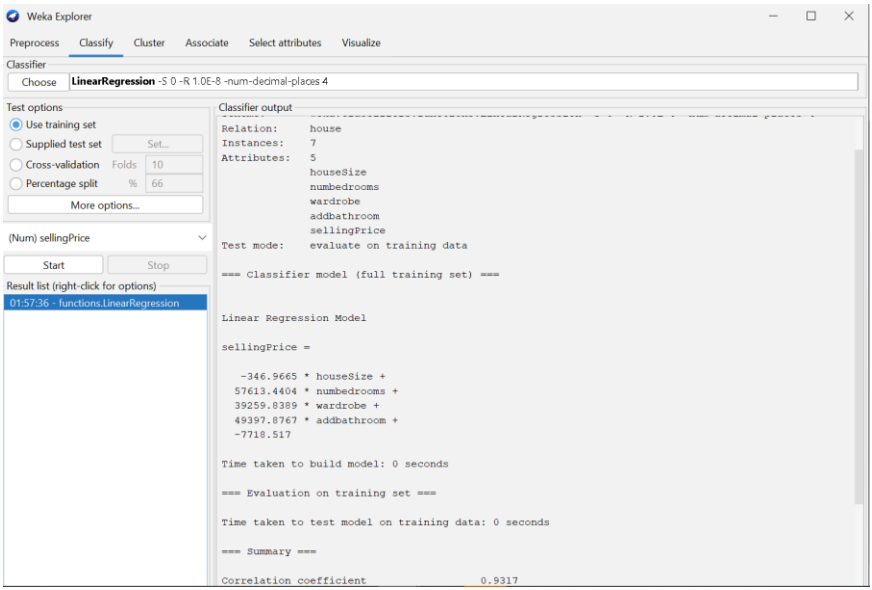


Рис. 4. Побудова регресійної моделі.

**Завдання**.

Побудувати регресійну модель та спрогнозувати ціну об’єкту ринку нерухомості (обрати за бажанням, наприклад, будинок) на основі даних сучасного ринку. Врахувати формування даних щодо об’єктів ринку нерухомості (50 об’єктів, 5 параметрів, формат файлу ARFF) та використання пакету WEKA. Зробити висновки.

**Питання для контролю рівня знань:**

1. В чому полягає метод регресійного аналізу?
2. Що визначає коефіцієнт кореляції ?
3. Який метод використовується при побудові рівняння регресії?

## Лабораторна робота №3

**Тема:** Інтелектуальний аналіз даних: метод класифікації з використанням програмного пакету WEKA.

**Теоретичні відомості:**

Ще одним методом інтелектуального аналізу даних, окрім методу регресійного аналізу, є метод класифікації. Метод класифікації – це метод аналізу даних, який дозволяє оцінити ймовірність того, що екземпляри даних належать до деякого класу залежно від значень їх атрибутів. При побудові моделі класифікації використовуються відомі значення атрибутів екземплярів даних та відповідно зв'язки між цими значеннями. При наявності нових екземплярів даних невідомого класу до даних застосовується раніше побудована модель класифікації і визначається відповідний клас.

Для створення моделі класифікації у **WEKA** потрібно підготувати файл з даними (формат **ARFF**), що буде використовуватися при дослідженні. Набір даних зазвичай ділять на дві частини так, щоб частина даних використовувалася для побудови моделі, тобто навчання, а інша частина використовуються для її перевірки, тобто коректності. Це дозволяє пересвідчитись, що модель працює не тільки під конкретний набір даних. Таким чином, варто розділити вибраний набір даних на два файли *.arff в співвідношенні 2/3 як навчальні дані та 1/3 як тестові від загальної кількості даних. Спочатку відкриваємо в програмному пакеті **WEKA** файл для навчання.

1) Скористаємося панеллю класифікації (**Classify**), рис. 1. Обираємо опцію **trees**, а потім опцію **J48**. Додатково потрібна опція **Use training set**, щоб пакет **WEKA** при створенні моделі класифікації використовував саме ті дані, які завантажили у вигляді ARFF-файлу. Натискаємо кнопку **Start**.

2) Щоб отримати дерево рішень, потрібно у контекстному меню обрати опцію **Visualize tree**. На екрані з'явиться зображення класифікаційного дерева нашої моделі (рис. 5), яке можна збільшити вибравши у контекстному меню опцію **Auto Scale.**

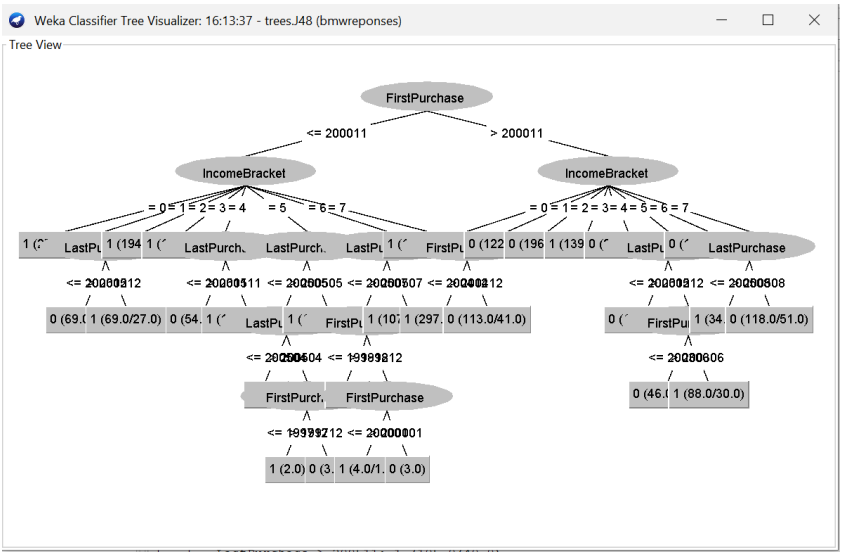

Рис. 5. Типове дерево рішень.

4) У вікні **Classifier output** отримаєте результати моделювання. Дерево рішень також розміщене під написом **J48 pruned tree**, тобто текстовий опис дерева з вузлами і листками. Найбільш суттєві дані – це показники класифікації: **Correctly Classified Instances,** який визначає точність побудованої моделі та **Incorrectly Classified Instances** – відповідно. Таблиця **Confusion Matrix** показує кількість хибно-позитивних і хибно-негативних розпізнавань.

3) Для перевірки класифікаційного дерева потрібно пропустити набір даних, що залишився (1/3 як тестові від загальної кількості даних) через отриману модель і перевірити, наскільки результати класифікації будуть відрізнятися від реальних даних. Обираємо у

секції **Test options** опцію **Supplied test set** та натискаємо на кнопку **Set.** Вказуємо тестовий файл, тобто 1/3 від загальної кількості, що містить решту даних, які не були включені в навчальний набір. Натискаємо кнопку **Start** і **WEKA** перевірить модель на основі тестових даних та покаже результат роботи моделі.

4) У вікні **Classifier output** отримаємо результати моделювання. Важливо, щоб показник класифікації **Correctly Classified Instances,** який визначає точність побудованої моделі, був близький до попереднього. Це означитиме, що нові дані, які будуть використовуватися в цій моделі в майбутньому, не знижуватимуть точність її роботи.

**Завдання.**

Побудувати модель класифікації, використати набір даних, зібраних дилерським центром BMW (розподіл за доходами; рік/місяць купівлі першого автомобіля BMW; рік/місяць купівлі останнього автомобіля BMW; інформація, чи скористався клієнт розширеною гарантією). Оскільки центр починає рекламну кампанію що передбачає розширену дворічну гарантію своїм постійним клієнтам. Подібні кампанії вже проводилися і дилерський центр має 4500 показники щодо попередніх продажів з розширеною гарантією.

Файли bmw-training.arff використати для навчання та bmw-test.arff для тестування (https://learnersdesk.weebly.com/weka-tutorials.html). Зробити висновки.

**Питання для контролю рівня знань:**

1. В чому полягає метод класифікації?
2. Як перевірити точність моделі класифікації ?
3. Чи потрібно формувати дані для перевірки коректності моделі?

## Лабораторна робота №4

**Тема:** Інтелектуальний аналіз даних: кластеризація з використанням програмного пакету WEKA.

**Теоретичні дані:**

Метод кластеризації дозволяє розбивати дані на групи за певною ознакою. Для розбиття множини даних на групи може використовуватися будь-який атрибут, а кількість груп визначається у методі кластеризації безпосередньо дослідником. Математичні методи, що використовуються у кластерному аналізі є складними, проте досить просто виконуються у програмному пакеті **WEKA**.

Для реалізації методу кластеризації даних у **WEKA** відкриваємо на закладці **Preprocess** ARFF-файл, що буде використовуватися при дослідженні.

1) Для розбиття даних на кластери обираємо закладку **Cluster**. При натисканні на кнопку **Choose** у запропонованому меню обираємо опцію **SimpleKMeans**, що є одним із методів кластеризації. Переконуємося, що обрана опція **Use training set**, щоб пакет **WEKA** при створенні моделі використовував дані, які завантажені у ARFF-файлі.
2) Клацаємо мишкою на опції **SimpleKMeans,** отримуємо поле атрибутів алгоритму, серед яких поле **numClusters** вказує на кількість кластерів для розбиття. Зокрема, це значення потрібно обрати ще до створення моделі, оскільки за замовчуванням у програмному пакеті визначено значення **2.** Натискаємо кнопку **OK** для збереження вибраних параметрів. Натискаємо кнопку **Start** і **WEKA** сформує результат відповідно до кількості вказаних кластерів.
3) Для візуального подання даних натискаємо правою кнопкою мишки в секції **Result List** закладки **Cluster**, де у контекстному меню обираємо опцію **Visualize Cluster Assignments** (рис. 6). В

результаті чого відкриється вікно з графічним представленням результатів кластеризації, де можна змінювати налаштування щодо осі **X** так і осі **Y**. Для кожного кластера передбачено виділення окремим кольором, опція **Color** в **Cluster (Nom)**, рис. 7.

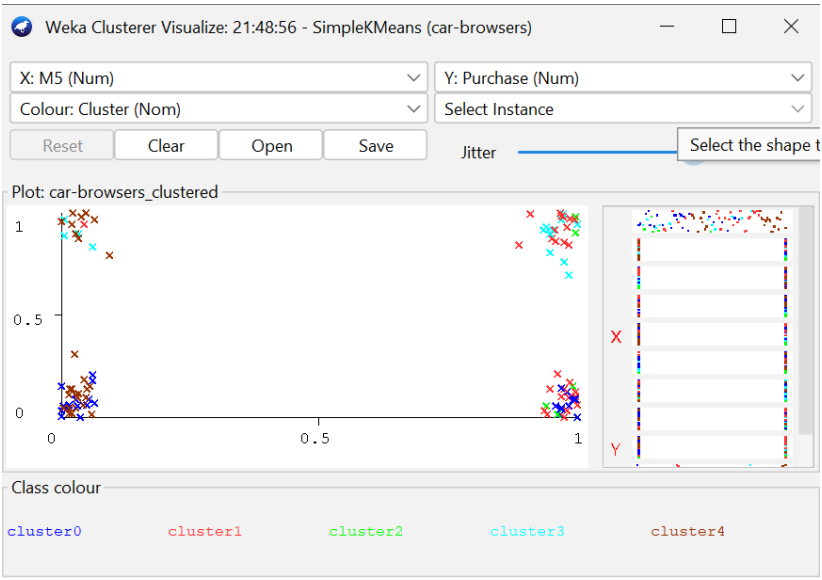


Рис. 6. Приклад реалізація опції
Visualize Cluster Assignments при кластерному аналізі.

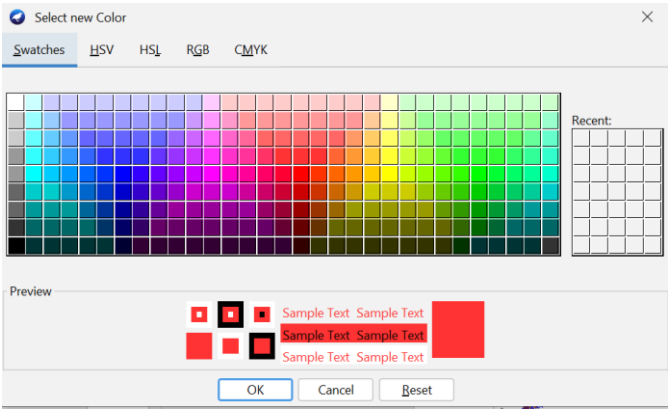


Рис. 7. Виділення кластера окремим кольором.

При цьому покажчик **Jitter** визначає величину розкиду між групами точок для зручного перегляду.

**Завдання.**

Побудувати модель кластеризації на основі даних файлу bmw-browsers.arff (https://learnersdesk.weebly.com/weka-tutorials.html) про відвідувачів демонстраційного залу щодо автомобілів, які їх зацікавили, та наскільки часто відвідувачі в підсумку купували автомобіль, який їм сподобався. Проаналізувати ці дані, щоб виділити різні групи відвідувачів та зрозуміти тенденції у поведінці відвідувачів. Зробити висновки.

**Питання для контролю рівня знань:**

1. В чому полягає метод кластеризації?
2. Як задавати кількість кластерів при реалізації методу?
3. Яка відмінність між методом класифікації та кластеризації ?

## Лабораторна робота №5

**Тема:** Інтелектуальний аналіз даних: метод найближчих сусідів за допомогою програмного пакета WEKA.

**Теоретичні відомості:**

Математичний алгоритм методу найближчих сусідів схожий до алгоритму методу кластеризації. Метод визначає відстань між невідомою точкою і всіма відомими точками даних, що є тривіальною задачею. Зокрема, найпоширеніший спосіб визначення відстані між двома точками – це нормалізована евклідова відстань (позначається $\|p-q\|$), що формально подається у вигляді:

$$\|p-q\| = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2} = \\ = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

де точки: $p=(p_1, p_2, ..., p_n)$ *та* $q=(q_1, q_2, ..., q_n)$.

Для реалізації методу найближчих сусіді у програмному пакеті **WEKA** відкриваємо на закладці **Preprocess** ARFF-файл, що буде використовуватися при дослідженні.

1) Скористаємося панеллю класифікації (**Classify**), рис.1, де обираємо опцію **lazy**, а потім **Ibk,** де ***IB*** **(Instance-Based)** – навчання на прикладах, ***k*** визначає кількість сусідів, поведінку яких ми хочемо дослідити. Зміна кількості найближчих сусідів в моделі визначається параметром **KNN** у вікні параметрів моделі, що з'являється при натисканні правою кнопки мишки на полі **IBk-K 1** ... . Точність моделі підвищується при додавання кількості сусідів.

2) Для побудови моделі за допомогою програмного пакету **WEKA** слід переконатися, що обрана опція **Use training set**, щоб використовувати набір даних, який завантажили початково до **WEKA**. Натискаємо кнопку **Start.**
3) У вікні **Classifier output** отримаємо результати моделювання. Зокрема, показник класифікації: **Correctly Classified Instances** визначатиме точність побудованої моделі методом найближчих сусідів та **Incorrectly Classified Instances** – відповідно. Таблиця **Confusion Matrix** показує кількість хибно-позитивних і хибно-негативних розпізнавань.

Використання методу найближчих сусідів потребує проведення великої кількості обчислень. Проте це не є проблемою для хмарних систем, оскільки обчислювальні процеси можуть ефективно виконуватися паралельно на багатьох комп'ютерах. Після проведення підрахунку всіх відстаней, результати будуть порівнюватися між собою для визначення найближчих сусідів. Але в більшості випадків задача спрощується, використовується лише частина бази даних та скорочується обсяг обчислень.

**Завдання.**

Побудувати модель методом найближчих сусідів на основі даних файлу bmw-training.arff (https://learnersdesk.weebly.com/weka-tutorials.html), що описаний у завданні лабораторної роботи №3. Зробити висновки.

**Питання для контролю рівня знань:**

1. В чому полягає метод найближчих сусідів?
2. Чи вливає кількість сусідів на точність методу?
3. Що є спільного між методом кластеризації та методом найближчих сусідів ?

## РЕКОМЕНДОВАНА ЛІТЕРАТУРА


1. Бахрушин В. Є. Методи аналізу даних : навчальний посібник для студентів. Запоріжжя : КПУ, 2011. 268 с.
2. Данильченко О. М., Данильченко А. О. Інтелектуальний аналіз даних : навч. посібник. Житомир : ЖДТУ, 2009. 405 с.
3. Іванов С. М., Максишко Н. К., Бречко Д. О. Інтелектуальний аналіз даних : конспект лекцій для здобувачів ступеня вищої освіти бакалавра спеціальності «Економіка» освітньо-професійної програми «Економічна кібернетика». Запоріжжя: ЗНУ, 2020. 156 с. URL: https://moodle.znu.edu.ua/pluginfile.php?file=/485849/mod_resource/content/1/%d0%91%d0%b0%d0%b7%d0%be%d0%b2%d0%b8%d0%b9%20%d0%bf%d1%96%d0%b4%d1%80%d1%83%d1%87%d0%bd%d0%b8%d0%ba%20%d0%b4%d0%b8%d1%81%d1%86%d0%b8%d0%bf%d0%bb%d1%96%d0%bd%d0%b8.pdf
4. Марченко О. О., Россада Т. В. Актуальні проблеми Data Mining : навч. посіб. Київ, 2017. 150 с.
5. Олійник А. О., Субботін С. О., Олійник О. О. Інтелектуальний аналіз даних : навч. посіб. Запоріжжя : ЗНТУ, 2012. 278 с.
6. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних : навч. посібник. К. : КНЕУ, 2007. 376 с.
7. Черняк О. І. Інтелектуальний аналіз даних : підручник. К. : Знання, 2014. 599 с.