Лабораторна робота №3

Інтелектуальний аналіз даних за допомогою програмного пакета WEKA

Що таке інтелектуальний аналіз даних?

За своєю суттю, інтелектуальний аналіз даних - це перетворення великих обсягів «сирих» даних в такі, що мають певний сенс і практично корисні схеми, структури і правила. Аналіз даних може бути розділений на два види - **прямий** (прогнозування) і **непрямий** (класифікація і кластеризація). Завдання ***прямого*** **аналізу** - прогноз конкретних показників, наприклад, прогноз продажної вартості будинку на базі інформації про ціни на будинки в даному районі.

Завдання ***непрямого*** **аналізу** - створення груп даних або пошук певних структур або схем в існуючому наборі даних. Наприклад, кожен перепис населення має на увазі інтелектуальний аналіз даних, так як уряд прагне отримати дані про кожного жителя і перетворити їх в інформацію, придатну для подальшого практичного використання.

Інтелектуальний аналіз даних в тому сенсі, в якому ми його розглядаємо виник в середині 90-х років минулого сторіччя, коли розвиток комп'ютерних технологій вийшов на досить високий рівень, а вартість обчислювальних потужностей і систем зберігання даних знизилася настільки, що компанії змогли дозволити собі самостійно проводити аналіз даних, не вдаючись до послуг великих обчислювальних центрів.

Крім того, слід зазначити, що термін «**інтелектуальний аналіз даних**», або **data mining**, є всеосяжним і включає в себе безліч різних підходів і методів для дослідження і перетворення даних.

Основна мета інтелектуального аналізу даних полягає в тому, щоб створити модель, що дозволяє ефективно інтерпретувати і використовувати ті дані, якими ви володієте зараз, і ті дані, які ви отримаєте в майбутньому. Оскільки аналіз даних включає в себе безліч методів, то основний етап створення моделі даних - це вибір методу аналізу, використовуваного в цій моделі. Для правильного вибору методу потрібен практичний досвід і деякі інструкції. Далі модель потрібно доопрацювати, щоб зробити її більш ефективною.

WEKA

Waikato Environment for Knowledge Analysis (WEKA), є вільно поширюваним програмним пакетом з відкритим вихідним кодом для аналізу даних.

Перша версія WEKA вийшла в 1993 році в університеті Ваїкато (Нова Зеландія). Вона була написана на різних мовах програмування. У 1997 році було прийнято рішення переписати програму на мову JAVA.

WEKA забезпечує графічний користувальницький інтерфейс для роботи з файлами даних і генерації візуальних результатів (у вигляді таблиць і графіків). Крім того, можливо інтегрувати WEKA, як і будь-яку іншу бібліотеку, у свої власні додатки, наприклад, для автоматизації аналізу даних на стороні сервера, використовуючи стандартний API.

WEKA поширюється по ліцензії GNU General Public License (GPL).

Ця програма дає можливість виконувати завдання аналізу даних як:

- підготовка даних - попередня обробка;
- відбір ознак;
- кластеризація;
- класифікація, зокрема, дерева рішень;
- пошук асоціативних правил;
- регресійний аналіз;
- візуалізація результатів;

Переваги WEKA

- об'ємний набір алгоритмів з аналізу даних і машинного навчання;
- відкритий вихідний код;
- кросплатформеність;

- простота у використанні;
- гнучкість у роботі з даними, що вводяться;
- вільний доступ;

Рис 1. Стартове вікно WEKA

При запуску WEKA, пакет пропонує вам на вибір 4 графічних інтерфейси для роботи з WEKA і даними. Будемо використовувати опцію **Explorer**. Її функціональності більш ніж достатньо для вирішення наших завдань.

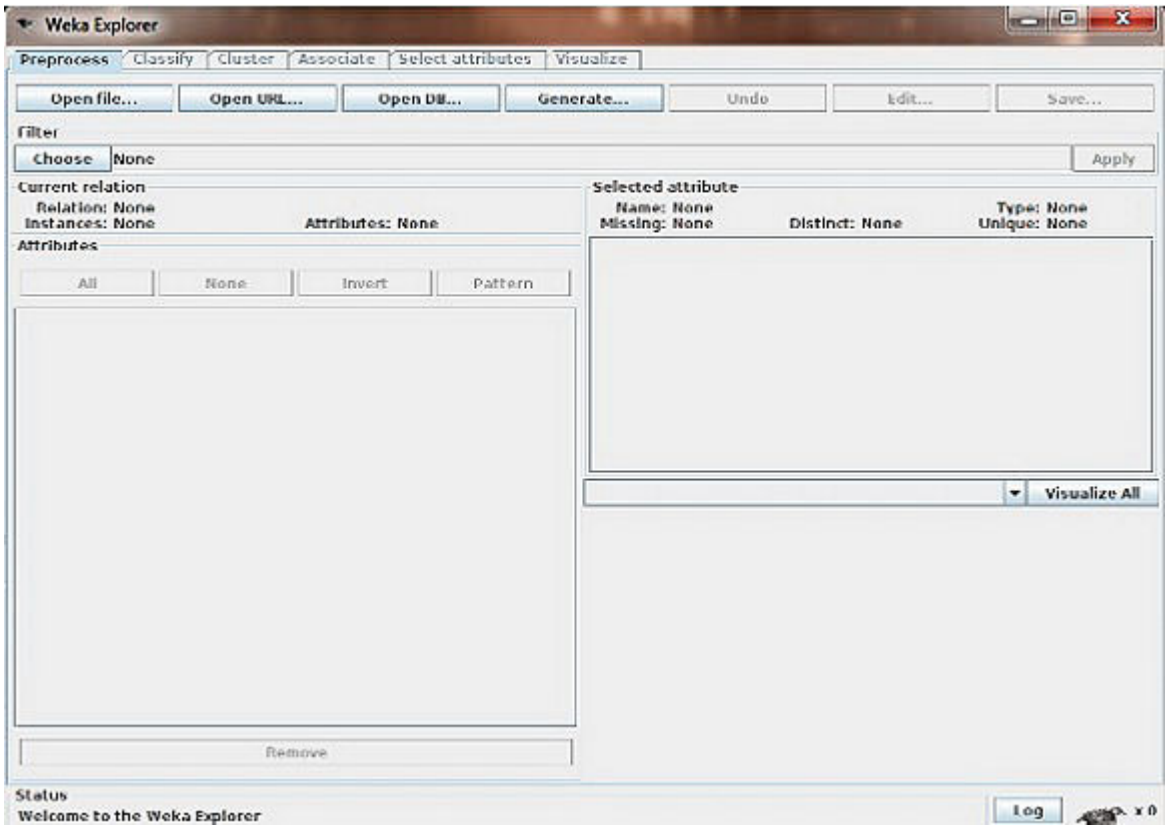


Рис 2. Вікно WEKA Explorer

Регресійний аналіз

Метод регресійного аналізу є найпростішим і, одним з найменш ефективнх методів інтелектуального аналізу даних. Найпростіша модель аналізу використовує один **вхідний** (**незалежний**) параметр і один **результуючий** (**залежний**) параметр (прикладами такої моделі є точкові діаграми Excel і аналогічні їм XYDiagram в OpenOffice.org). Безумовно, модель можна ускладнити, додавши кілька десятків вхідних параметрів, але в будь-якому

випадку загальний підхід буде один і той же: на підставі кількох незалежних змінних визначається один залежний результат. Таким чином, модель регресійного аналізу використовується для прогнозування значення однієї залежної змінної, виходячи з відомих значень декількох незалежних параметрів.

Найбільш очевидний приклад регресійного аналізу - визначення вартості будинку. Ціна на будинок (залежна змінна) визначається декількома незалежними параметрами: яка площа будинку і розмір ділянки, чи використовуються в оформленні кухні гранітні плити, яка якість і термін служби сантехніки і так далі.

Скористаємося моделлю регресійного аналізу для визначення ціни будинку і розберемо конкретний приклад. У таблиці внизу вказані фактичні параметри будинків, виставлених на продаж в певному районі. На підставі цих даних спробуємо оцінити вартість будинку в останньому рядку таблиці.

Таблиця 1. Регресійна модель оцінки вартості будинку

Площа будинку (кв.футів)	Розмір ділянки	Кількість спалень	Гранітна обробка на кухні	Сучасне сантехнічне обладнання?	Ціна продажу
3529	9191	6	0	0	$ 205,000
3247	10061	5	1	1	$ 224,900
4032	10150	5	0	1	$ 197,900
2397	14156	4	1	0	$ 189,900
2200	9600	4	0	1 `	$ 195,000
3536	19994	6	1	1	$ 325,000
2983	9365	5	0	1	$ 230,000
3198	***9669***	***5***	***1***	***1***	***????***

Розглянута нами модель дає лише найзагальніше, досить поверхневе, уявлення про метод регресійного аналізу. Тим не менш, нашого поверхневого розгляду цілком достатньо для того, щоб зрозуміти основні принципи і створити модель регресійного аналізу за допомогою WEKA.

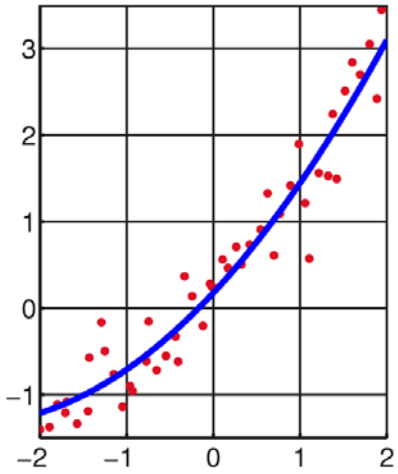
Розглянемо наступні поняття:

- метод найменших квадратів,

- нормальний розподіл,

- коефіцієнт детермінації R-квадрат

Метод найменших квадратів *(МНК, OLS, Ordinary Least Squares) - математичний метод, який застосовується для вирішення різних задач, заснований на мінімізації суми квадратів деяких функцій від шуканих змінних. Він може використовуватися зокрема для апроксимації точкових значень деякою функцією. МНК є одним з базових методів регресійного аналізу для оцінки невідомих параметрів регресійних моделей за вибірковими даними.*



Результат підгонки сукупності спостережень квадратичною функцією.

Нормальний розподіл (закон Гаусса)

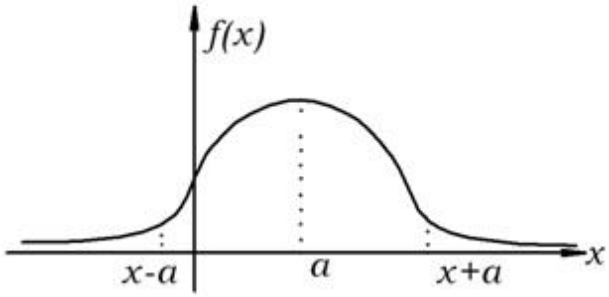
Нормальний закон розподілу (normal law of distribution) *(який ще називається законом Гаусса) відіграє виключно важливу роль в теорії ймовірностей і займає серед інших законів розподілу особливий стан. Це закон, який найчастіше зустрічається на практиці.*

Більшість випадкових величин, таких, наприклад, як похибки вимірів, похибки гарматних стрільб і т. д. можуть бути подані як суми великої кількості малих доданків - елементарних похибок, кожна з яких визначається дією окремої причини, яка не залежить від інших. Яким би законам розподілу не підпорядковувались окремі елементарні похибки, особливості цих розподілів в сумі великої кількості доданків нівелюються і сума підпорядковується закону, що близький до нормального. Підсумовані похибки в загальній сумі повинні грати відносно малу роль.

Випадкова величина ξ нормально розподілена або підпорядковується закону розподілу Гаусса, якщо її щільність розподілу має вигляд:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}},$$

де a - довільне дійсне число, $\sigma > 0$.



Коефіцієнт детермінації *(позначається як R^2 — R-квадрат) — статистичний показник, що використовується в статистичних моделях як міра залежності варіації залежної змінної від варіації незалежних змінних. Вказує наскільки отримані спостереження підтверджують модель.*

Створення набору даних для завантаження в WEKA

Для того щоб завантажити дані в WEKA, їх слід перетворити у формат, зрозумілий для цього програмного пакету. Найбільш підходящим форматом для завантаження даних в WEKA є формат Attribute-Relation File Format (ARFF), який спочатку визначає тип завантажуваних даних, а потім вказує власне дані. У файлі формату ARFF ви вказуєте назву і тип даних для кожного стовпця таблиці, а потім дані по рядках. У моделях регресійного аналізу використовуються всього два типи даних: **NUMERIC і DATE**. Після того, як ви

описали всі стовпці таблиці, ви додаєте дані по рядках, використовуючи як роздільник кому. Нижче наведено файл ARFF з даними про ціни на будинки, які ми будемо використовувати для побудови нашої тестової моделі. Зверніть увагу, що в списку відсутній рядок з даними будинку, ціну для якого необхідно встановити. Зараз ми створюємо регресійну модель на базі відомих параметрів і, отже, не можемо включити в неї параметри нашого будинку, оскільки ціна його невідома.

Файл даних для завантаження в WEKA

```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

Завантаження даних в WEKA

Тепер, коли файл з даними готовий, його потрібно завантажити в WEKA. Запустіть WEKA і виберіть опцію **Explorer**. В результаті відкриється закладка **Preprocess** вікна Explorer. Клацніть на кнопці **Open File** і виберіть створений вами ARFF-файл. Вікно WEKA Explorer із завантаженими даними про будинках показано на рис. 3.

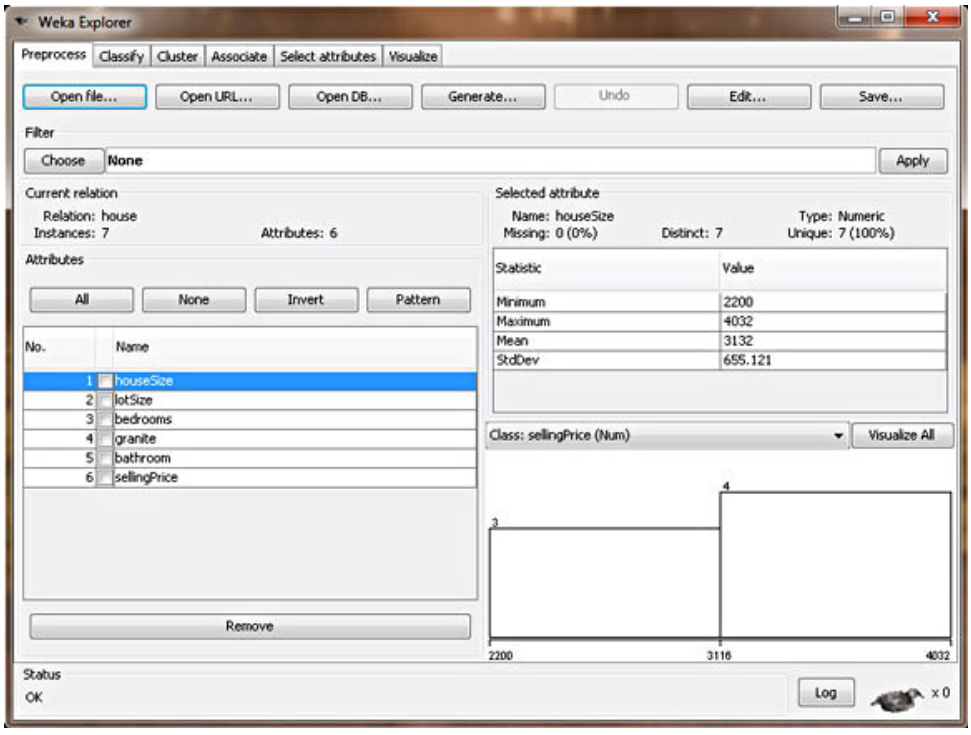


Рис. 3. Вікно WEKA Explorer із завантаженими даними про будинки

У цьому вікні ви можете перевірити дані, на підставі яких ви збираєтеся будувати модель. У лівій частині вікна **Explorer** показані параметри об'єктів (**Attributes**), які відповідають заголовкам стовпців нашої вихідної таблиці, а також вказано кількість об'єктів (**Instances**), тобто рядків таблиці. Якщо ви клацнете мишкою на одному із заголовків стовпців, то в правій панелі буде виведена повна інформація про набір даних в даному стовпці. Наприклад, якщо ми виберемо стовпець **houseSize** в лівій панелі (він обраний за замовчуванням), то в правій панелі відобразиться додаткова статистична інформація з цього стовпця. Буде показано максимальне значення в стовпці (4032 кв.футів) і мінімальне значення (2200 кв.футів). Крім того, буде підраховано середнє значення (3131 кв.фут) і стандартне відхилення (655 кв.футів) (стандартне відхилення - статистичний показник розсіювання значень випадкової величини). Нарешті, тут же вам пропонується можливість візуального аналізу даних (кнопка **Visualize All**). Оскільки в нашій таблиці даних не так багато, то їх візуальне відображення не дає такої наочної аналітичної картини, як у випадку використання сотень або тисяч показників.

Давайте перейдемо від розгляду даних до створення моделі і визначимо, нарешті, вартість мого будинку.

Створення регресійної моделі в WEKA

Для того щоб створити модель, відкрийте закладку **Classify**. В якості першого кроку, нам треба вибрати тип моделі для аналізу, щоб вказати WEKA, яким чином ми хочемо аналізувати наші дані, і яку модель побудувати:

1. Клацніть на копанні **Choose** і розгорніть меню **functions**.
2. Виберіть опцію **LinearRegression** .

Таким чином, ми вказали WEKA, що ми хочемо створити модель регресійного аналізу. Як ви помітили, меню включає безліч моделей. Це ще раз підкреслює факт нашого дуже поверхневого знайомства з областю інтелектуального аналізу даних. Зверніть увагу: у меню включена опція **SimpleLinearRegression** , проте ми не використовуємо її, оскільки цей тип моделі визначає значення залежної змінної за значеннями одного незалежного параметра, а у нас їх цілих шість. Якщо ви вибрали правильну модель, то вікно WEKA Explorer має виглядати так, як показано на рис. 4.

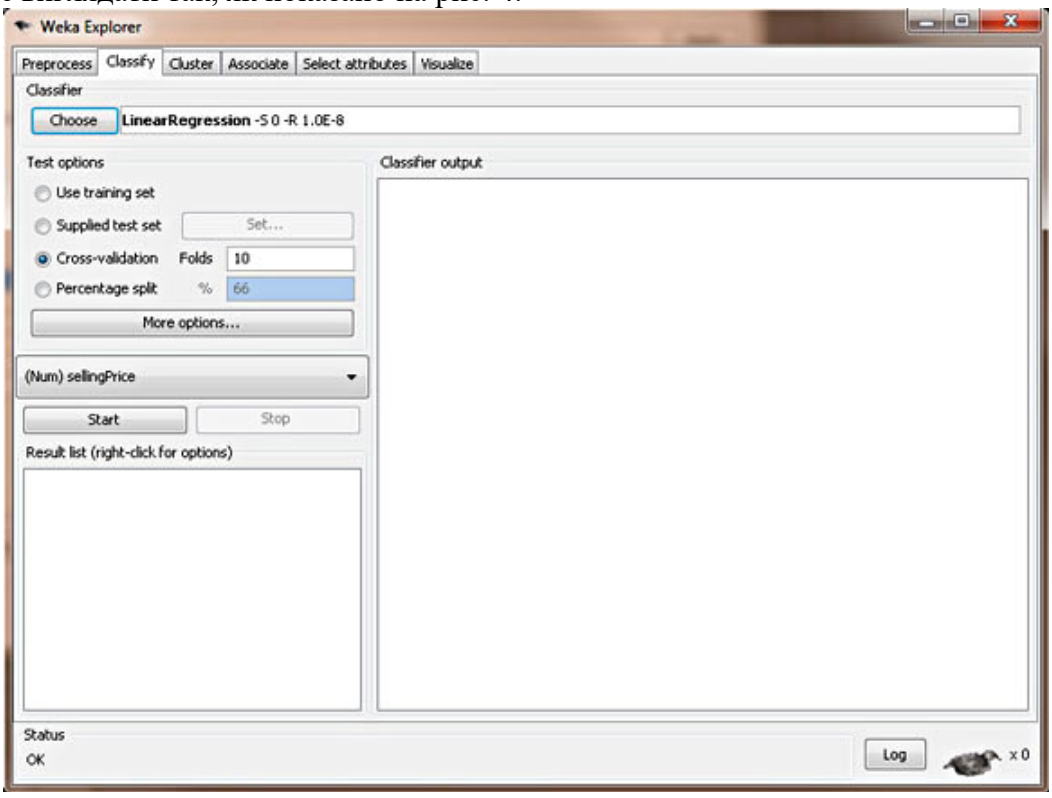


Рис. 4. Модель лінійного регресійного аналізу WEKA

Після того, як ми вибрали тип моделі, потрібно вказати WEKA, які дані повинні використовуватися для її створення. Незважаючи на те, що відповідь на це питання для нас

цілком очевидна - потрібно взяти дані зі створеного нами ARFF-файлу - існує кілька інших, більш складних, можливостей надання даних для аналізу. Опція **Supplied test set** дозволяє вказати додатковий набір тестових даних для моделі, опція **Cross-validation** використовує кілька наборів даних, усереднює їх і будує модель на основі середніх значень, а опція **Percentage split** використовує в якості бази для моделі процентилі набору даних. Ці способи застосовуються для створення аналітичних моделей. У разі регресійного аналізу нам потрібна опція **Use training set** . У цьому випадку WEKA створить модель на базі даних із завантаженого ARFF-файлу.

Завершальний етап створення моделі - вибір залежної змінної (колонка, в якій знаходиться невідоме нам значення, яке потрібно розрахувати). У нашому прикладі - це ціна будинку, оскільки, саме це значення ми і хочемо дізнатися. Відразу після секції **Test options** знаходиться список, що розкривається, в якому вам потрібно вибрати залежний параметр. Типово повинен бути вибраний атрибут **sellingPrice** . Якщо це не так, виберіть самі цей параметр.

Ми визначили всі параметри і можемо приступити до створення моделі. Натисніть кнопку **Start** . У результаті вікно WEKA має виглядати так, як показано на рис. 5.

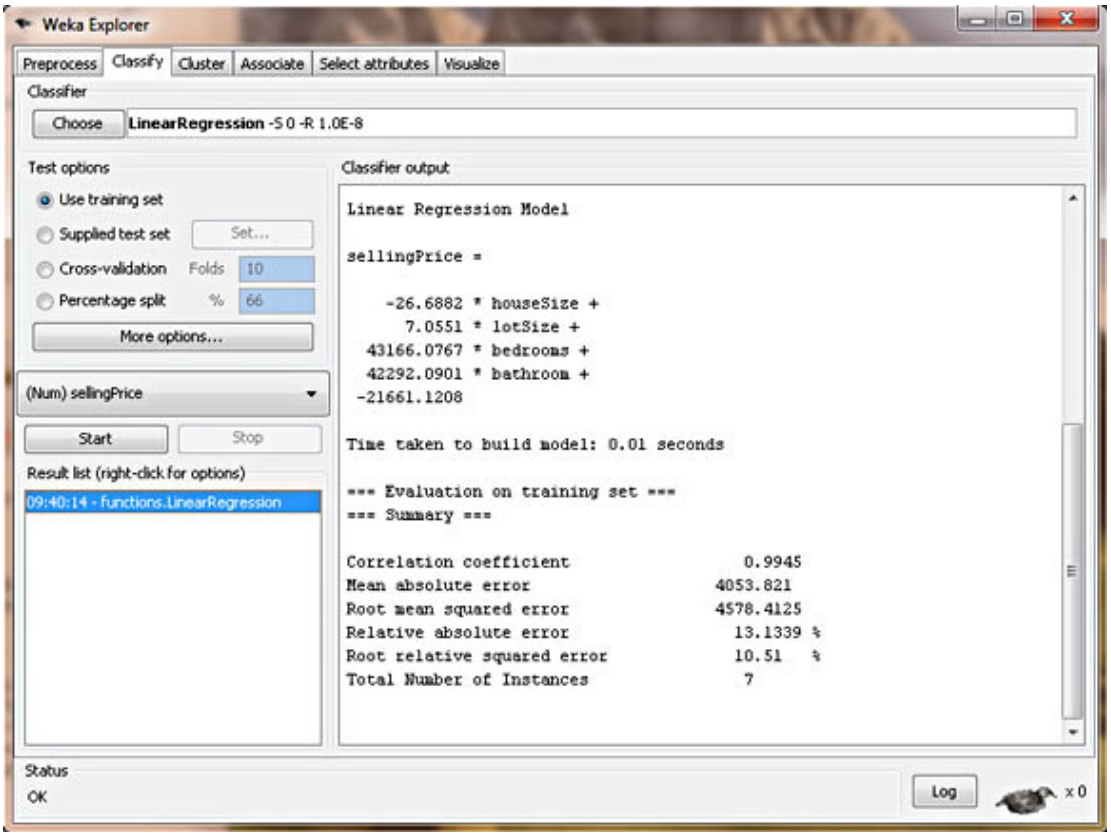


Рис. 5. Регресійна модель WEKA для розрахунку вартості будинку

Інтерпретація результатів регресійного аналізу

Розберемо, які дані включені в результуючий висновок

Готова модель регресійного аналізу

```
sellingPrice = (-26.6882 * houseSize) +
               (7.0551 * lotSize) +
               (43166.0767 * bedrooms) +
               (42292.0901 * bathroom)
              - 21661.1208
```

Далі в отриману модель для визначення вартості підставлені параметри нашого будинку.

Розрахунок вартості будинку на базі готової моделі

```
sellingPrice = (-26.6882 * 3198) +
               (7.0551 * 9669) +
               (43166.0767 * 5) +
               (42292.0901 * 1)
              - 21661.1208
sellingPrice = 219,328
```

Однак, можливості інтелектуального аналізу даних не обмежуються визначенням одного параметра. Основне завдання аналізу - виявлення залежностей і зв'язків у великих наборах даних. Інтелектуальний аналіз, як правило, використовується не для того, щоб визначити яке-небудь конкретне значення, а для того, щоб побудувати модель, що дозволяє аналізувати зв'язки між даними, прогнозувати результати і робити обґрунтовані висновки, які підтверджуються зібраними статистичними даними. Тому не будемо обмежуватися розрахованою ціною будинку: розглянемо залежності між даними нашої моделі і постараємося зробити певні висновки щодо правил формування цін на нерухомість.

- **Гранітні елементи в оформленні кухні не впливають на ціну будинку** - WEKA використовує тільки ті дані, які, згідно зі статистикою, впливають на точність моделі (вплив кожного незалежного параметра на залежну змінну визначається за допомогою коефіцієнта детермінації). Таким чином, параметри, що не мають достатнього впливу на залежну змінну, в моделі не враховуються. Наша регресійна модель свідчить про те, що використання граніту на кухні не впливає на ціну будинку.
- **Стан ванних кімнат та сантехніки впливає на ціну будинку** - оскільки ми використовуємо значення 0 або 1 в якості показника модернізації ванних кімнат, то відповідний коефіцієнт в регресійній моделі демонструє нам, як сучасне сантехнічне обладнання впливає на ціну будинку, а саме додає 42292 $ до його ціни.
- **Велика площа будинку знижує його ціну** - Відповідно до моделі WEKA, у міру зростання площі будинків, ціна знижується. Це випливає з того, що модель включає змінну **houseSize** з негативним коефіцієнтом. Що ж виходить? Збільшення площі будинку на 1 кв.фут знижує його вартість на 26$? Подібне твердження здається очевидною нісенітницею. Насправді, розмір будинку не є незалежною величиною. Цей параметр пов'язаний, наприклад, з кількістю спалень - очевидно, що у великих будинках і кількість спалень більше. Так що наша модель, на жаль, не ідеальна, але ми можемо її поправити. Закладка **Preprocess** дозволяє видалити стовпці з набору даних.

В якості самостійної вправи, видаліть стовбець **houseSize** і створіть нову модель. Перевірте, як зміна набору даних відіб'ється на ціні будинку, і яка з двох моделей більше відповідає реальності (уточнена ціна будинку $ 217,894).

Розглянемо більш реальний приклад. Для створення моделі скористаємося файлом даних, пропонованим в якості бази для регресійного аналізу на Web-сайті проекту WEKA. Теоретично, новий приклад буде дещо складнішим нашої примітивної моделі, що використовує дані про сім будинків. Пропонований файл призначений для створення регресійній моделі розрахунку витрати бензину (MPG - кількості миль на галон), виходячи з декількох параметрів автомобіля (дані збиралися з 1970 по 1982 рік). Модель враховує декілька параметрів машини - кількість циліндрів, робочий об'єм двигуна, його потужність, вага автомобіля, час розгону, рік випуску, виробника і марку автомобіля. Цей набір даних містить 398 рядків і відповідає більшості вимог до статистичних даних, чого не можна сказати про наш попередній набір даних про будинки. Теоретично, модель на основі нового набору даних буде значно складнішою, і WEKA доведеться докласти більших зусиль на розробку нової моделі.

Для побудови моделі регресійного аналізу на основі нового набору даних вам слід виконати всі ті ж кроки, що і для моделі аналізу ціни будинку, так що ми не будемо приводити їх повторно. Висновок, який повинен вийти в результаті регресійного аналізу, показаний далі:

Модель регресійного аналізу для визначення MPG

```
class (aka MPG) =

-2.2744 * Cylinders = 6,3,5,4 +
-4.4421 * Cylinders = 3,5,4 +
 6.74 * cylinders = 5,4 +
 0.012 * displacement +
-0.0359 * Horsepower +
-0.0056 * Weight +
 1.6184 * model = 75,71,76,74,77,78,79,81,82,80 +
 1.8307 * model = 77,78,79,81,82,80 +
 1.8958 * model = 79,81,82,80 +
 1.7754 * model = 81,82,80 +
 1.167 * model = 82,80 +
 1.2522 * model = 80 +
 2.1363 * origin = 2,3 +
 37.9165
```

З точки зору виконання обчислень, створення потужних регресійних моделей на базі великих масивів даних, не викликає особливих проблем. Модель для визначення MPG може здатися набагато складнішою, ніж модель для визначення вартості будинку, тим не менш, це не так. Наприклад, перший рядок моделі, -2.2744 * cylinders = 6,3,5,4 означає, що якщо у машини 6-цілінрового двигун, то потрібно в формулу підставити 1, а якщо 8-циліндровий двигун - то 0. Давайте підставимо в модель реальні дані (наприклад, з рядка 10) і перевіримо, наскільки результат обчислень буде відповідати реальному показнику.

Обчислення показника MPG

```
      data = 8,390,190,3850,8.5,70,1,15
class (aka MPG) =
    -2.2744 * 0 +
    -4.4421 * 0 +
     6.74 * 0 +
     0.012 * 390 +
    -0.0359 * 190 +
    -0.0056 * 3850 +
     1.6184 * 0 +
     1.8307 * 0 +
     1.8958 * 0 +
     1.7754 * 0 +
     1.167 * 0 +
     1.2522 * 0 +
     2.1363 * 0 +
    37.9165

Expected Value = 15 mpg
Regression Model Output = 14.2 mpg
```

Таким чином, при використанні випадково вибраних даних, результат роботи нашої моделі (14.2 MPG) виявився досить близькою до реального показника (15 MPG).

Контрольні запитання

1. Що таке WEKA?
2. Метод регресійного аналізу.
3. Метод найменших квадратів.
4. Закон нормального розподілу.