

література

Міністерство освіти і науки України
Тернопільський національний технічний університет ім. Івана Пулюя
Кафедра комп'ютерних наук



Навчально-методична

Методичні вказівки
до лабораторної роботи №4 з курсу
«Інтелектуальний аналіз даних»
для студентів спеціальності 122 «Комп'ютерні науки»

Тема: «Знаходження рівняння множинної регресії»

Тернопіль – 2017

Знаходження рівняння множинної регресії. Методичні вказівки до лабораторної роботи з навчальної дисципліни «Інтелектуальний аналіз даних», для студентів спеціальності 122 “Комп’ютерні науки”, кафедра КН ТНТУ ім. І. Пулюя, Тернопіль, 2017 р.

Укладач:

*ст. викл. кафедри КН ТНТУ ім. І. Пулюя **Козбур Галина Володимирівна***

Відповідальний за випуск:

*ст. викл. кафедри КН ТНТУ ім. І. Пулюя **Козбур Галина Володимирівна***

Лабораторна робота № 4

Тема: Знаходження рівняння множинної регресії

Мета роботи: Освоєння методу отримання рівняння множинної (багатофакторної) регресії.

Завдання: використовуючи програмний пакет WEKA, навчитись знаходити аналітичну форму залежності результативної ознаки від декількох факторних ознак (рівняння множинної регресії, або модель зв'язку).

Порядок виконання роботи

1. Встановлення WEKA та побудова рівняння лінійної регресії на основі навчальних даних.

Встановити програмний пакет WEKA (Waikato Environment for Knowledge Analysis). WEKA є програмним пакетом з відкритим вихідним кодом для аналізу даних. WEKA поширюється по ліцензії GNU General Public License (GPL). Одне із можливих джерел - <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>.



Рис 1. Стартове вікно WEKA

При запуску пакет пропонує на вибір кілька графічних інтерфейсів для роботи з WEKA і даними. Для вирішення нашого завдання будемо використовувати опцію **Explorer**.

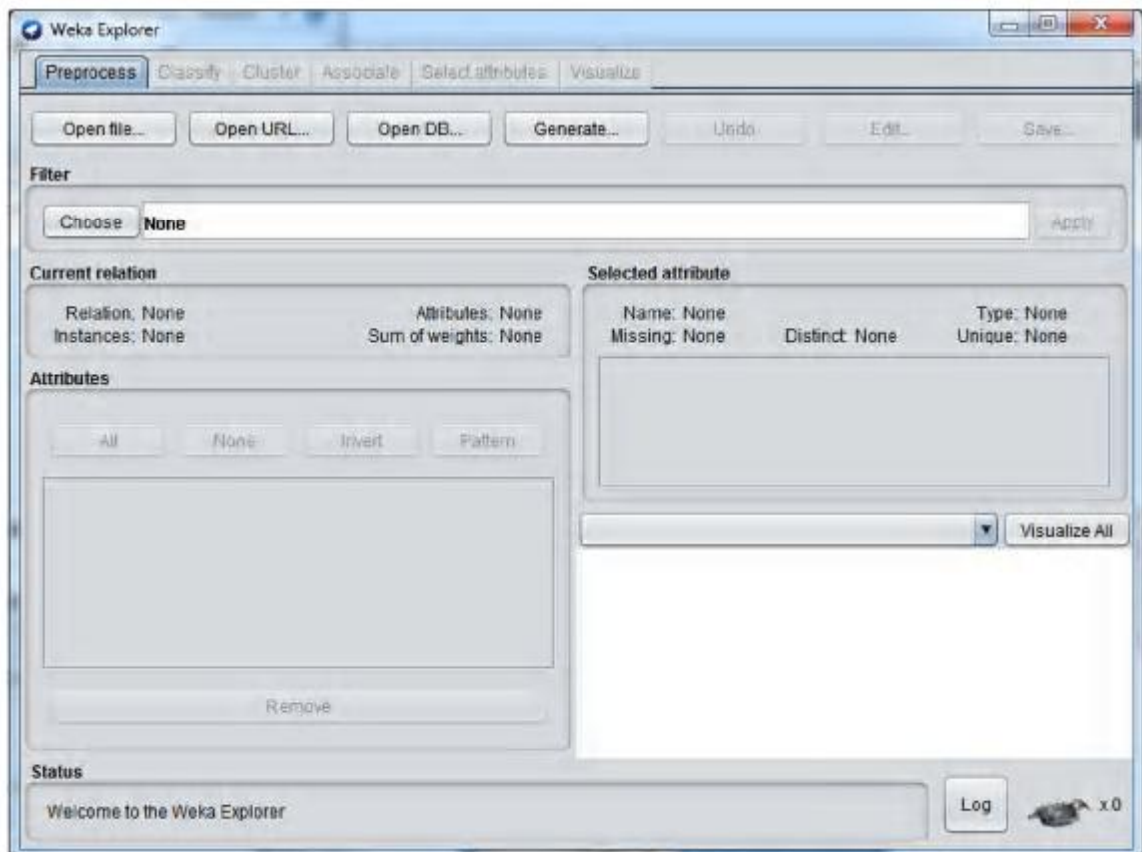
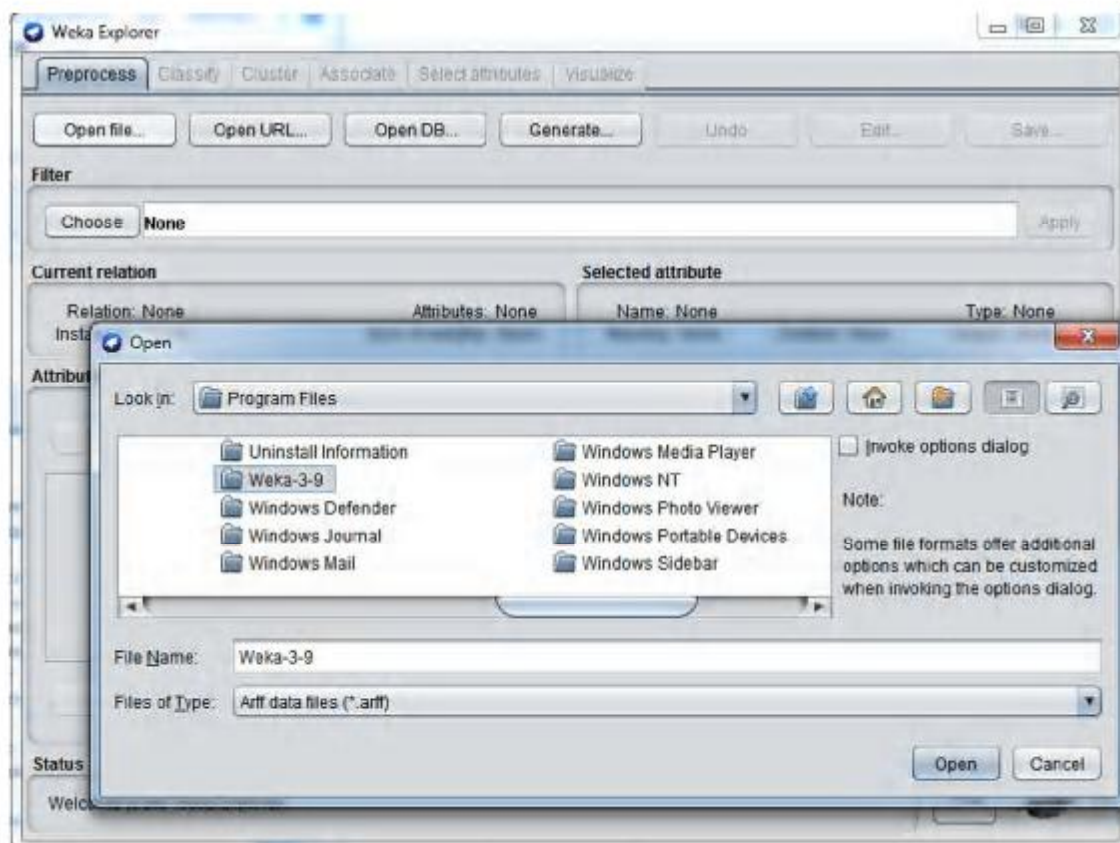
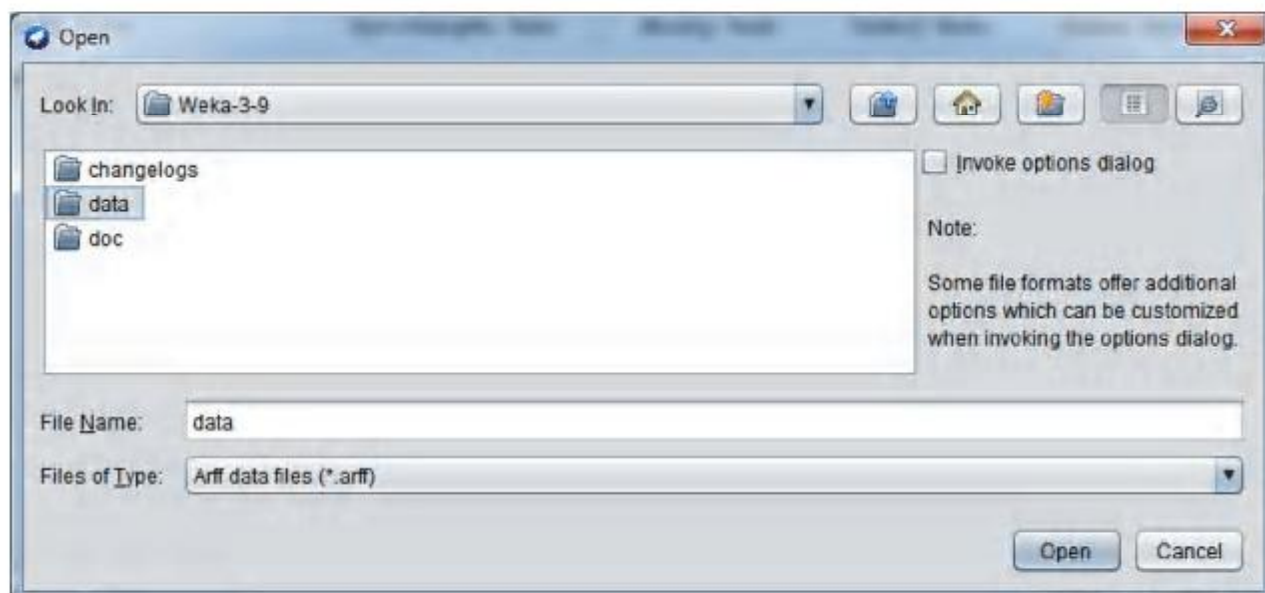


Рис 2. Вікно WEKA Explorer

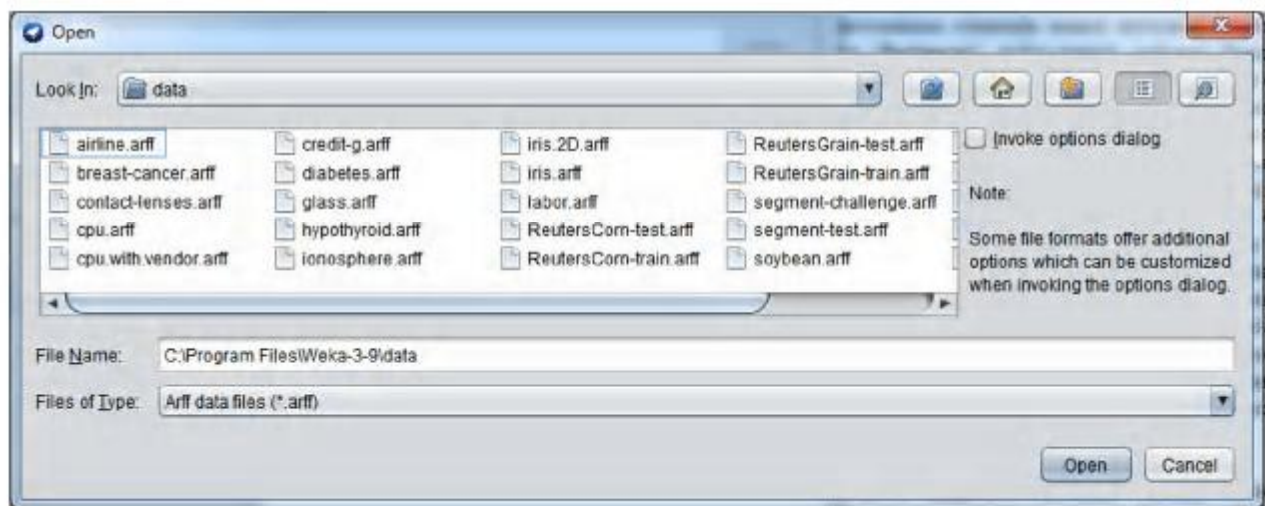
Після вибору опції **Open File...** вікно **WEKA Explorer** набуде вигляду:



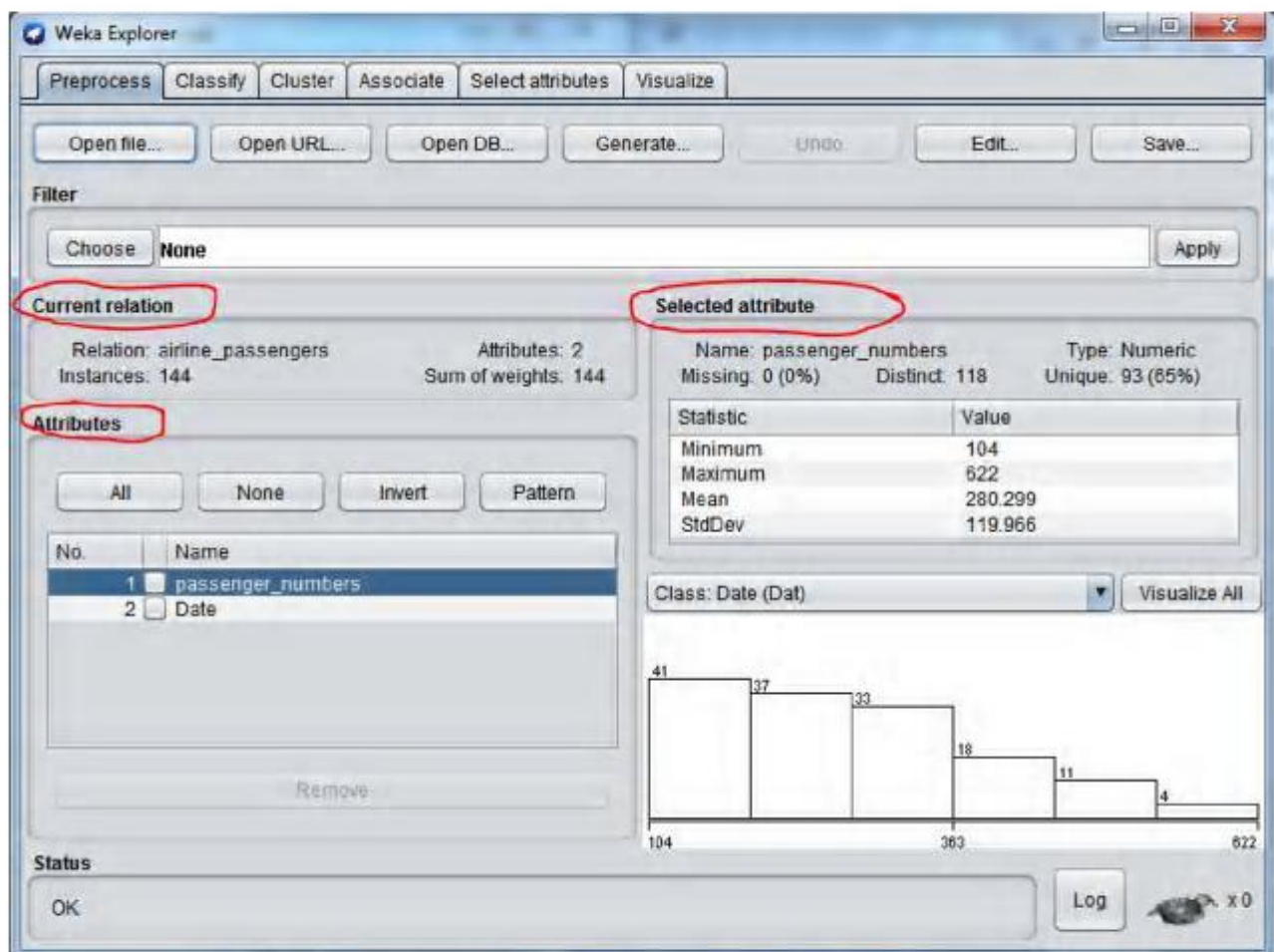
В папці файлів **WEKA-3-9** вибираємо папку з даними для навчальних цілей.



Відкриваємо файл `airline.arff`:



Вікно із завантаженими даними набуде вигляду:

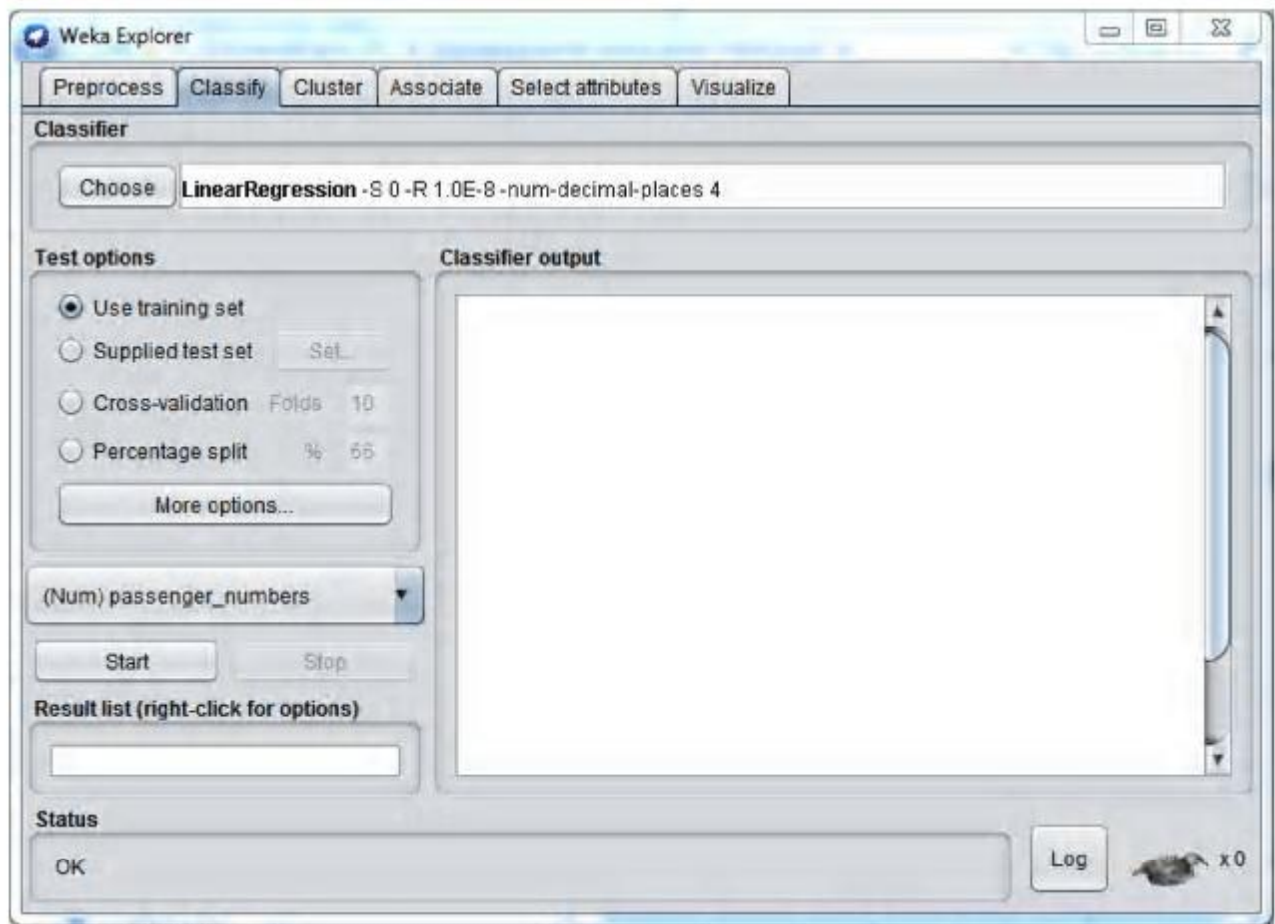


У лівій частині вікна показано дані, за якими потрібно збудувати модель: параметри об'єктів (**Attributes**), кількість таких атрибутів (2), а також вказано кількість об'єктів (**Instances**). Вибравши один із атрибутів, у правій панелі можна

побачити повну інформацію про набір числових даних по цьому атрибуту. Наприклад, при виборі атрибуту **passenger_numbers** в лівій панелі у правій панелі **Selected attribute** відобразиться додаткова статистична інформація з цього атрибуту: максимальне і мінімальне значення, середнє значення та середнє квадратичне відхилення. Є можливість візуального аналізу даних (**Visualize All**) з допомогою стовпчикових діаграм.

Як відомо з математичної статистики, найпростіша модель регресійного аналізу використовує один **вхідний (незалежний)** параметр і один **результуючий (залежний)** параметр, результатом є аналітичний вираз для функції однієї змінної $y = f(x)$, де x - незалежний параметр (фактор), y - залежний параметр (результат). У випадку багатьох вхідних параметрів визначається один залежний результат як функція багатьох змінних $y = f(x_1, x_2, \dots, x_n)$. Тут $x_1, x_2, \dots, x_n$ - незалежні аргументи функції, тобто атрибути, відібрані для побудови моделі, n - їх кількість. Отримана модель регресійного аналізу використовується для прогнозування значення залежної змінної, виходячи з відомих значень одного чи декількох незалежних параметрів (атрибутів).

Оскільки в нашій задачі використовується лише два атрибути, то один з них буде незалежним, або фактором (**passenger_numbers**), інший – залежним, або результатом. Побудуємо регресійну модель для навчальних даних. Для цього вибираємо тип задачі **Classify**, з допомогою **Choose** вибираємо з папки **functions** вид аналізу даних **LinearRegression**. Для навчальних даних вибираємо **Use training set**, вказуємо атрибут, який повинен бути результуючим: **(Num) passenger_numbers** та натискаємо **Start**.



У вікні **Classifier output** отримуємо результати обробки даних:

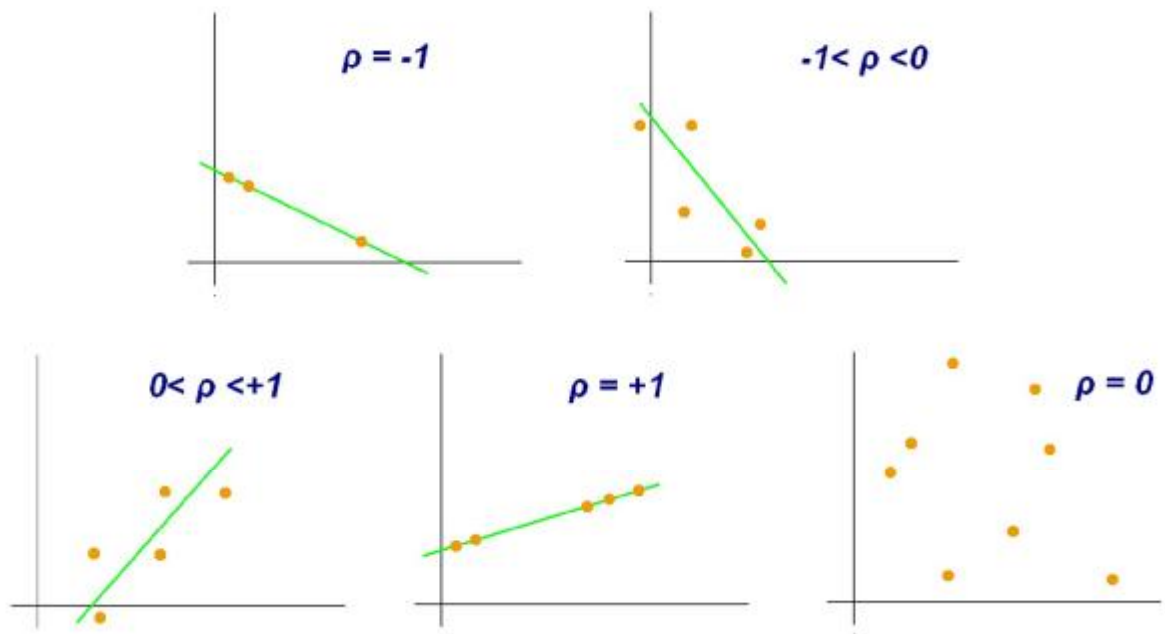
- **Run information** – інформація про вхідні дані.
- **Linear Regression Model** – вираз для шуканої функції $f(x)$, у нас $f(x) = 844774835.7902 \cdot x_{(\text{passenger_numbers})} - 711546413173.6233$. Після отримання цієї формули можна прогнозувати значення функції $f(x)$, підставляючи те чи інше значення аргументу x . Також подається час побудови моделі.
- **Evaluation on training set** – час тестування моделі.
- **Summary** – оцінка отриманої моделі. Коефіцієнт кореляції показує, наскільки сильно результативний атрибут залежить від факторного атрибуту (тіснота зв'язку). Якщо значення коефіцієнта кореляції лежить в проміжку $[0,8; 1]$, то отриманий зв'язок вважають дуже тісним.

Кореляція	Негативна	Позитивна
Відсутня	-0.09 до 0.0	0.0 до 0.09
Низька	-0.3 до -0.1	0.1 до 0.3
Середня	-0.5 до -0.3	0.3 до 0.5
Висока	-1.0 до -0.5	0.5 до 1.0

Додатна, або позитивна кореляція — кореляція, при якій збільшення однієї змінної пов'язане зі збільшенням іншої, при цьому коефіцієнт кореляції додатний. Від'ємна, або негативна кореляція — кореляція, при якій збільшення однієї змінної пов'язане зі зменшенням іншої, при цьому коефіцієнт кореляції від'ємний.

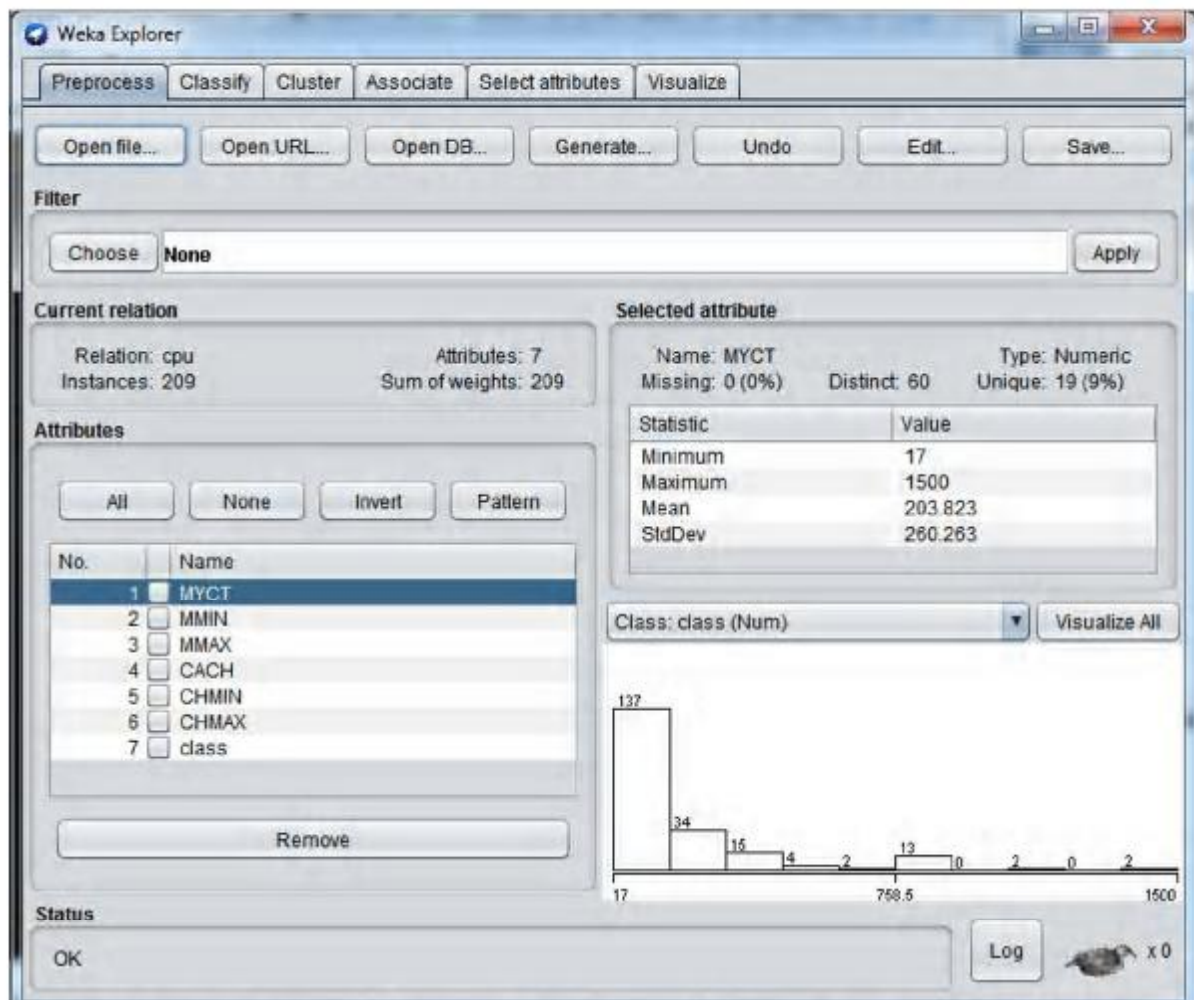
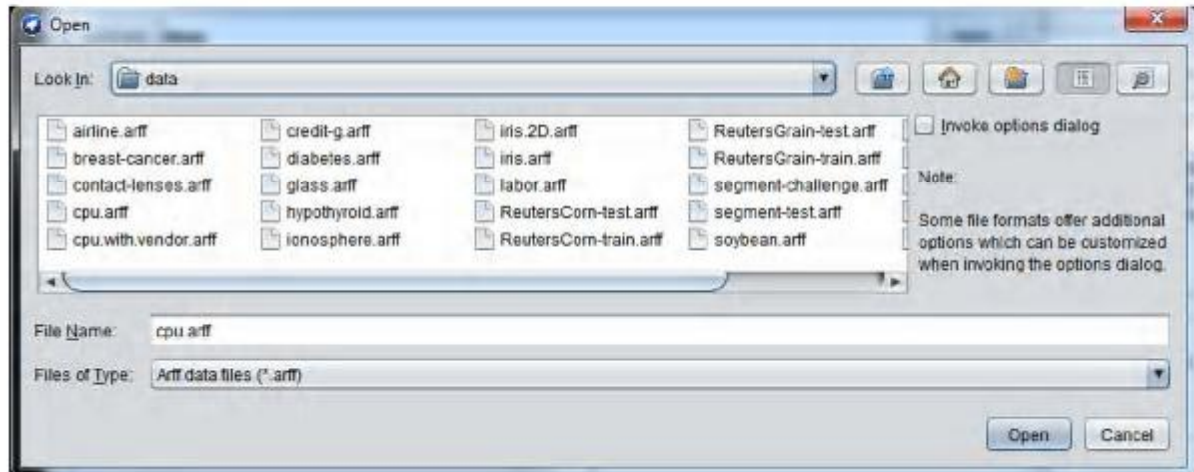
Також у пункті **Summary** отримуємо значення абсолютної та відносної похибок результату, в абсолютних одиницях та відсотках відповідно.

Візуалізацію зв'язку між атрибутами можна отримати з допомогою інструменту **Visualize**. Зразки діаграм розсіювання даних та відповідні коефіцієнти кореляції представлено на рисунку:



2. Побудова моделі множинної регресії на навчальних даних.

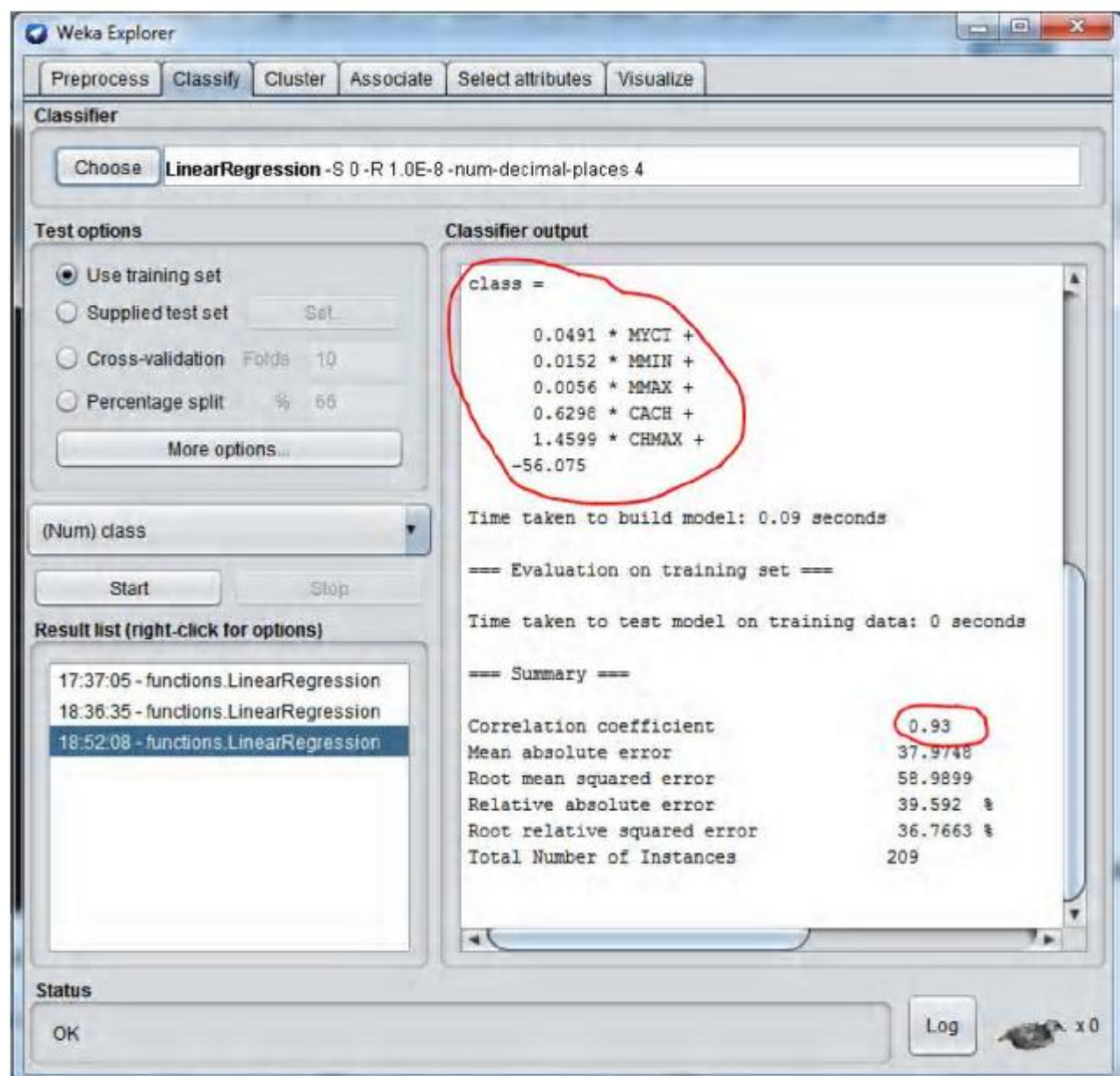
Вибравши в якості навчального файлу cpu.arff та повторивши усі дії, отримаємо рівняння множинної лінійної регресії. Відповідні скріншоти надано на рисунках.



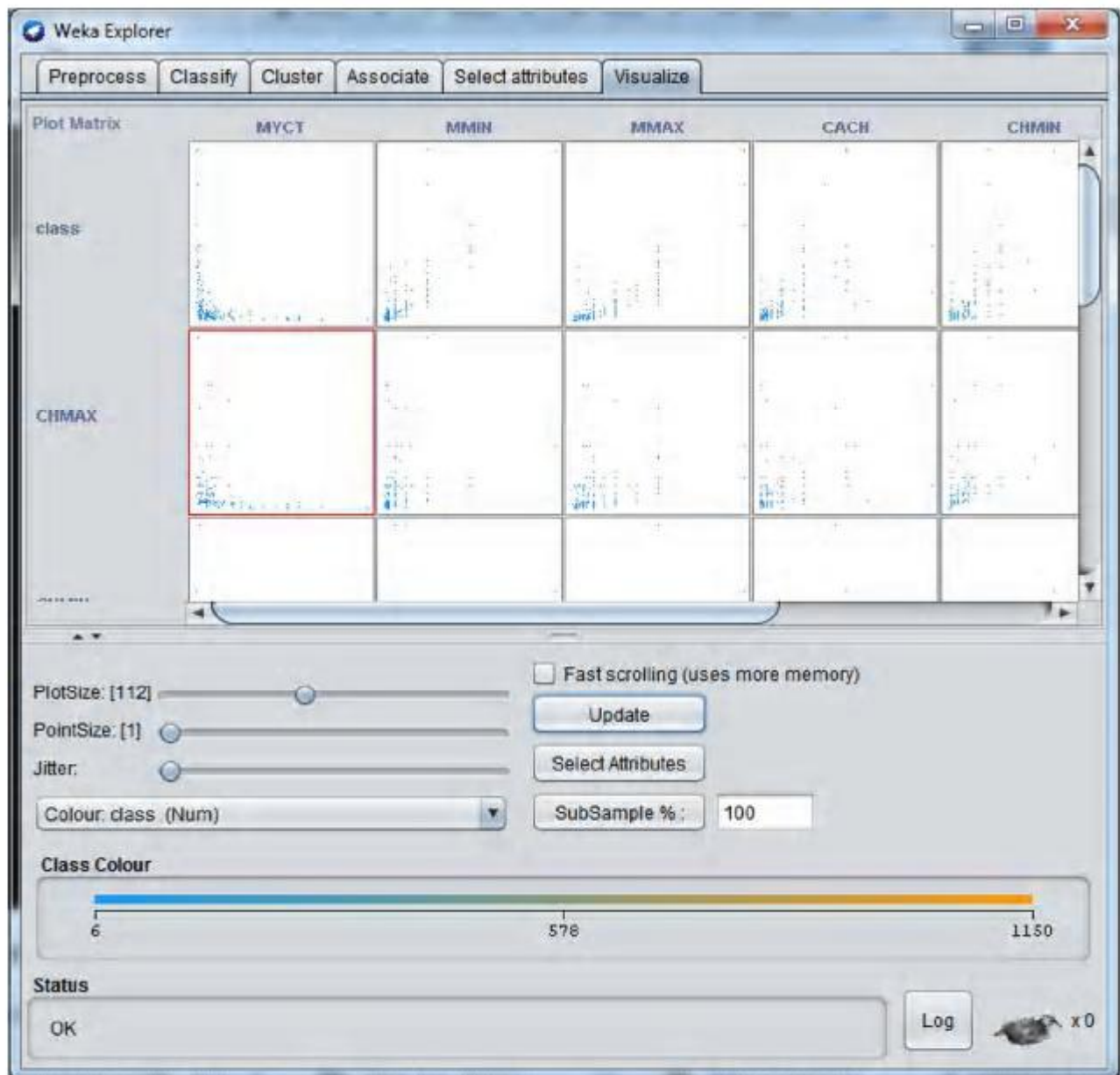
Серед 7 атрибутів один вибирається як результативний (**class**), інші 6 – факторні атрибути. Рядків з даними – 209. За цими даними отримано таку модель регресії:

$$f(x)_{class} = 0.0491 \cdot MYCT + 0.0152 \cdot MMIN + 0.0056 \cdot MMAX + 0.6298 \cdot CACH + 1.4599 \cdot CHMAX - 56.075.$$

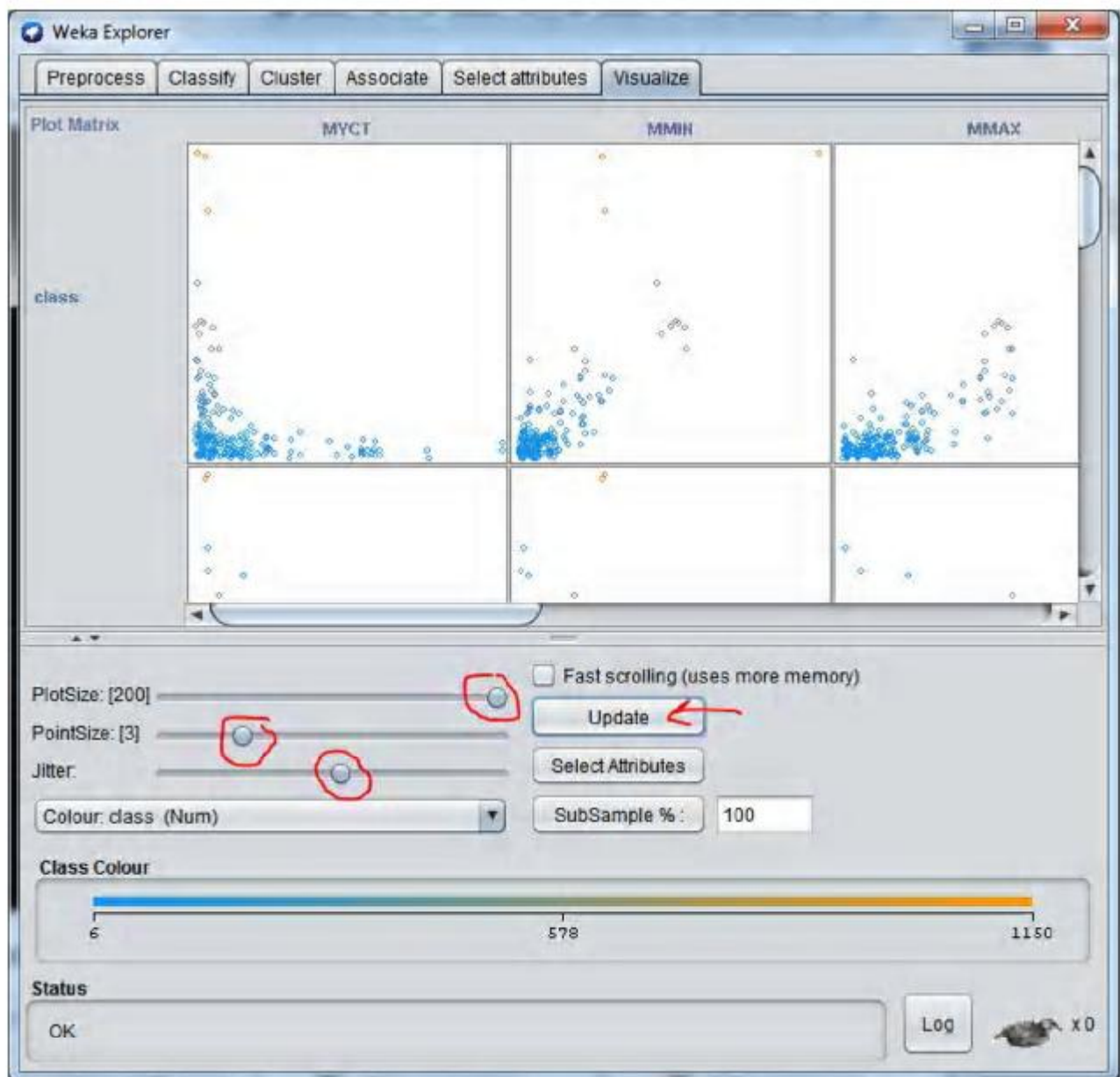
Підставляючи у формулу різні набори значень аргументів (MYCT, MMIN, MMAX, CACH, CHMAX), будемо отримувати відповідні (прогнознi) значення атрибуту **class**.



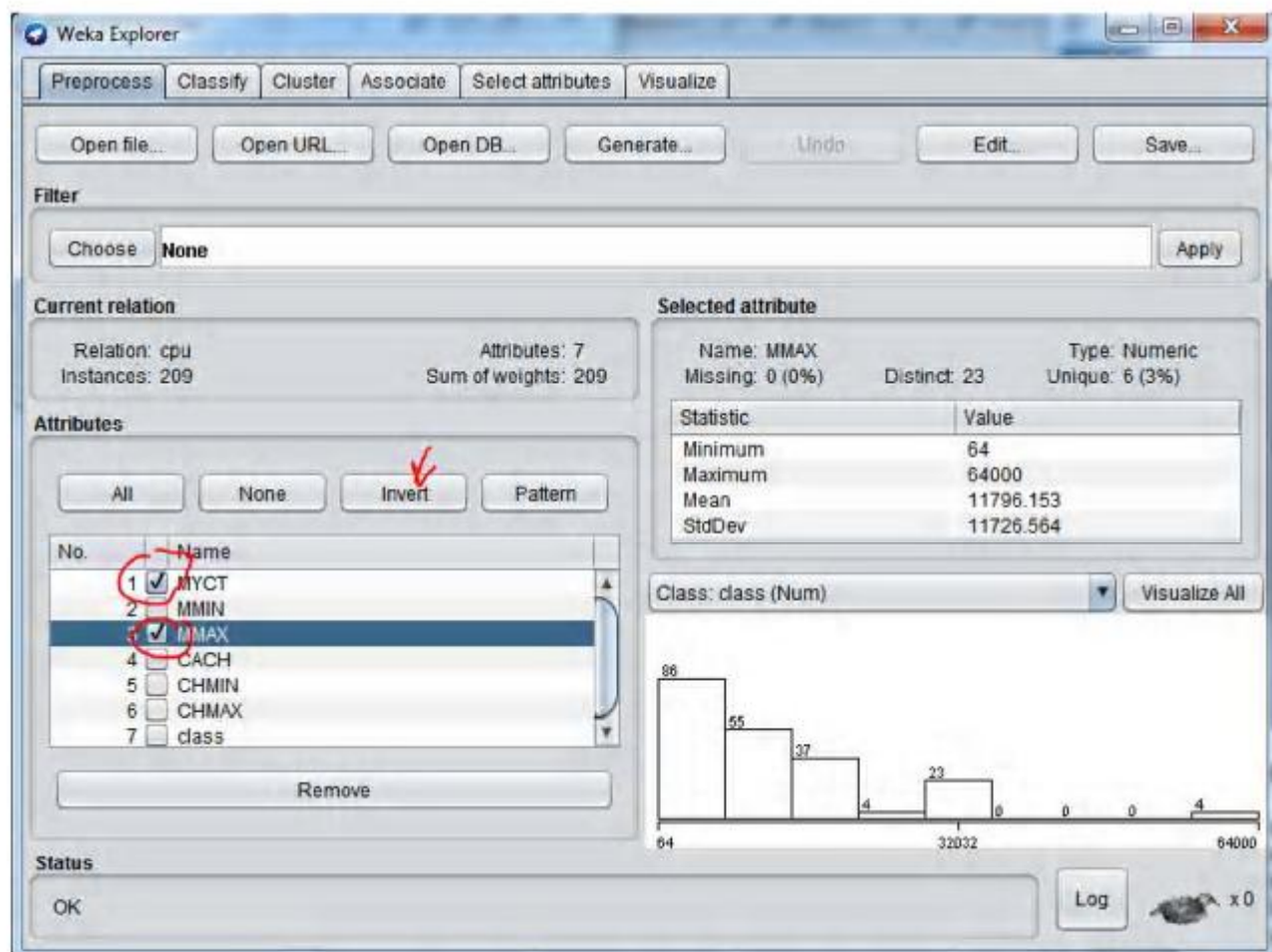
Використавши засіб візуалізації, отримуємо в єдиній таблиці графічне представлення попарних залежностей між атрибутами задачі.



Змінюючи параметри відображення, отримуємо картинку, зручну для візуального аналізу:



Зауважимо, що великі значення похибок або дуже малі значення коефіцієнтів у рівнянні регресії можуть свідчити про некоректність побудованої моделі. Причиною некоректності часто буває явище **мультиколінеарності**, - коли між незалежними, як вважається, атрибутами вже існує кореляційний зв'язок. Покращити модель можна, виключивши з числа факторних атрибутів менш важливі з числа тих, між якими прослідковується мультиколінеарність. Знайти мультиколінеарні атрибути можна, знайшовши ті пари незалежних атрибутів, для яких значення коефіцієнта кореляції є високим (>0.8).



В будь-якому разі побудові моделі повинен передувати загальний аналіз поставленої задачі, вхідних даних та очікуваного результату.

3. Підготовка власних даних та знаходження рівняння множинної регресії з аналізом результатів.

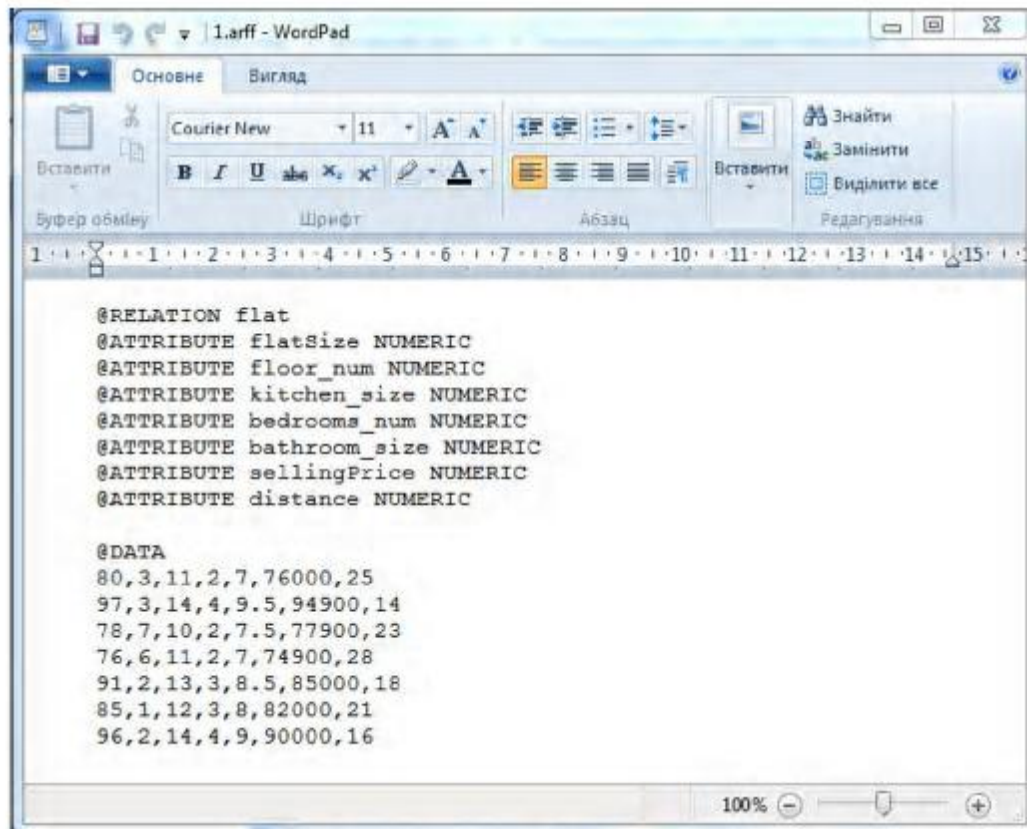
Розглянемо задачу: є база даних з продажу нерухомості в регіоні. Потрібно виконати оцінку ринкової ціни своєї квартири (будинку, дачної ділянки, тощо).

Для завантаження даних у WEKA їх слід перетворити у формат Attribute-Relation File Format (ARFF). Зауважимо, що у моделях регресійного аналізу використовуються всього два типи даних: **NUMERIC** і **DATE**.

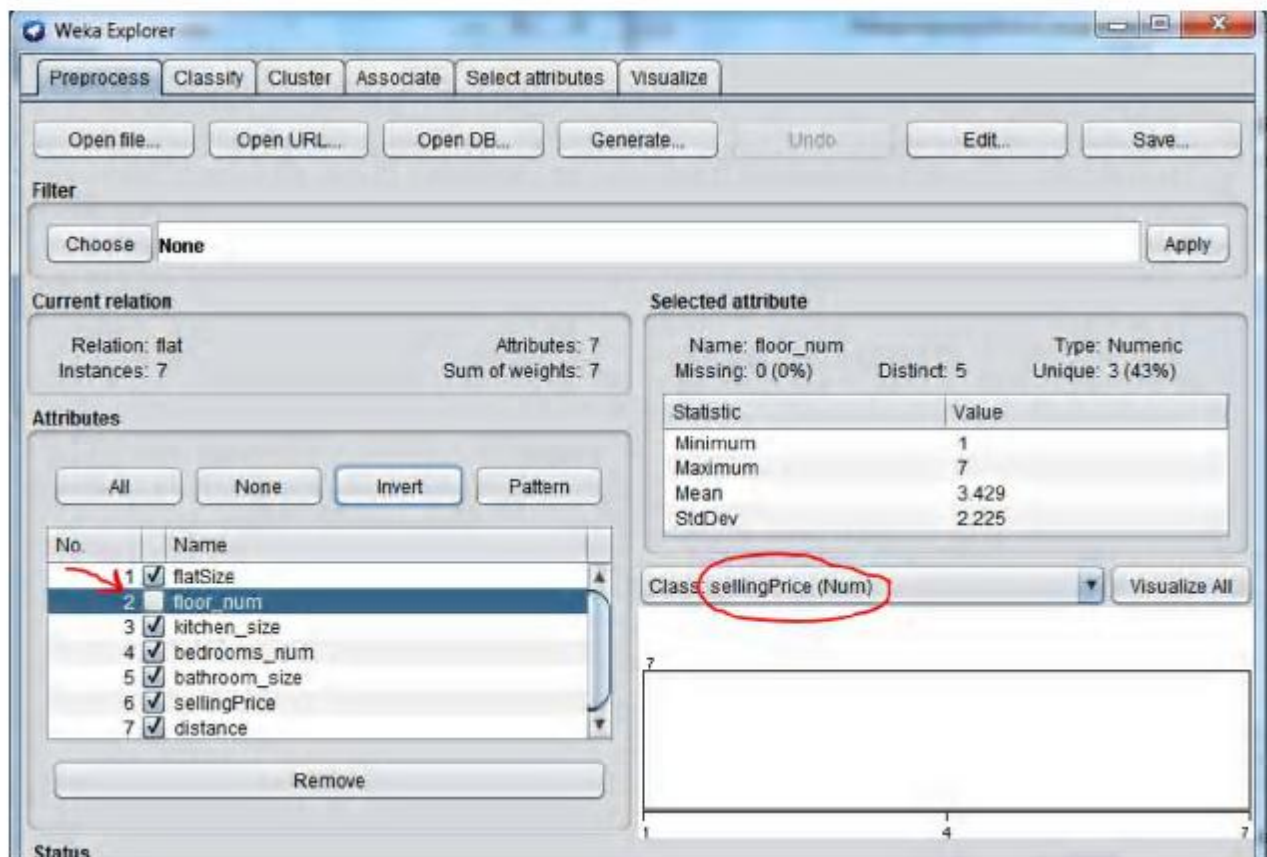
У файлі формату ARFF вказують назву і тип даних для кожного з атрибутів (наприклад, загальна площа, поверх, площа кухні, кількість спалень, площа санвузла, ціна продажу та відстань (у хв ходьби) від центру), а потім числові значення атрибутів для кожного об'єкта (наприклад, для першої квартири, взятої з відкритої бази ринку нерухомості, загальна площа становить 80 кв.м, площа кухні – 12 кв.м, кількість спалень – 2, поверх – 3, площа санвузла – 8 кв.м, вартість – 56 тис.дол., до центру населеного пункту – 25 хв пішки). Числові значення атрибутів для кожного об'єкту (запису) слід вказати, використовуючи як роздільник кому.

Нижче наведено файл ARFF з даними про ціни на будинки, які будуть використовуватись для побудови нашої моделі. Дані квартири (будинку, дачної ділянки), ціну для якої необхідно встановити, у таблиці відсутні, оскільки регресійна модель створюється на базі відомих параметрів, а ринкова ціна нашої квартири невідома.

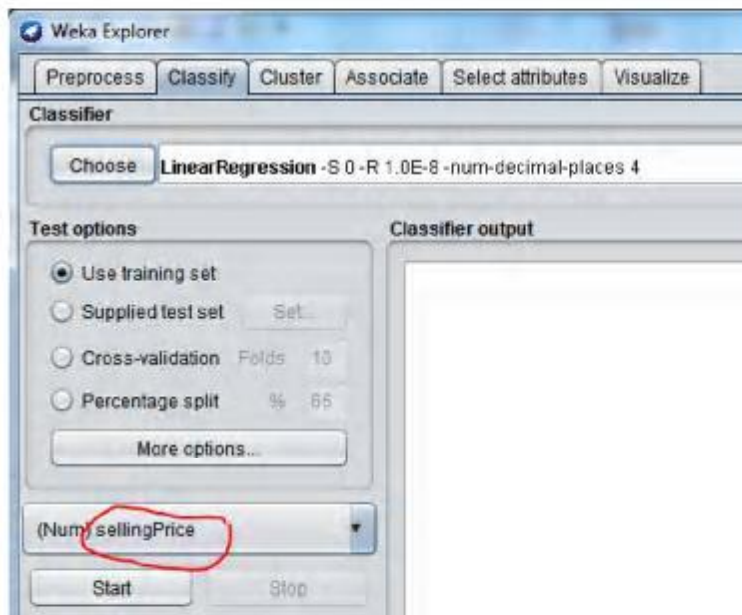
Файл даних для завантаження в WEKA

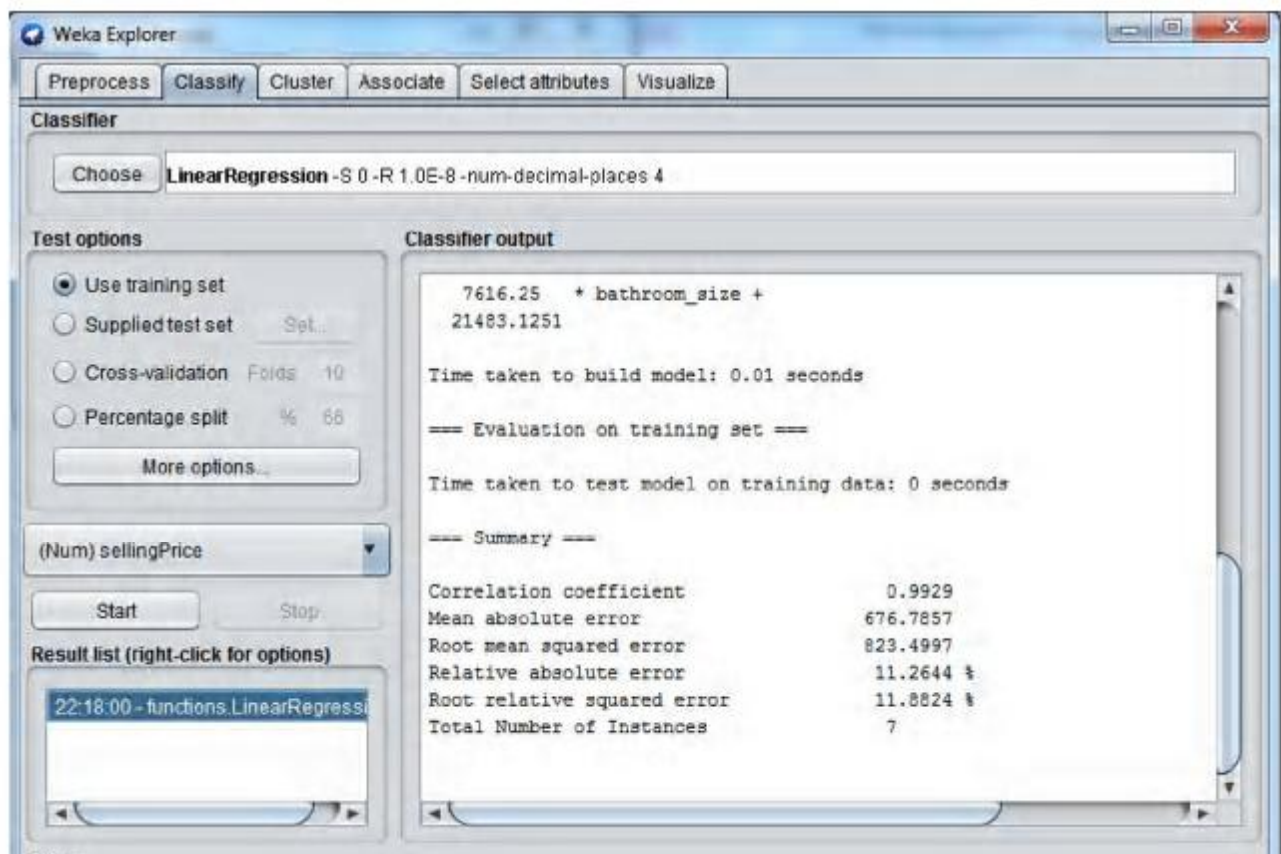


ARFF файл завантажують в WEKA з допомогою закладки **Preprocess** вікна Explorer.



У правому полі вказуємо результативний атрибут (**sellingPrice**). При побудові регресійної моделі з таким вхідним файлом отримуємо коефіцієнт регресії 0. Переглянувши з допомогою інструменту Visualize графічні представлення попарних залежностей між атрибутами задачі, бачимо, що поверх ніяк не впливає на вартість квартири (точки на діаграмі **floor_num-sellingPrice** розташовані безсистемно, хаотично). Тому виключаємо атрибут **floor_num** з розгляду. Отримуємо:





Визначити вартість своєї квартири можна, підставивши в готову формулу відповідні параметри житла.

4. 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 слід обмежуватися розрахованою ціною будинку, а зробити висновки щодо правил формування цін на нерухомість.

Зміст звіту по роботі:

1. Титульний лист.
2. Мета роботи.
3. Результати роботи з навчальними даними в пакеті WEKA з поясненнями та скріншотами.
4. Створення **власної бази даних** (не менше 30 записів із 4-6 атрибутами) для знаходження ринкової вартості об'єкту нерухомості, одиниці техніки чи транспортного засобу (описати об'єкт та зробити скріни з джерел(-а) інформації).
5. Знаходження рівняння регресії або встановлення неможливості побудови моделі.
6. Аналіз результатів.
7. Висновки по роботі.

Індивідуальні завдання лабораторної роботи №4:

Варіант 1. Оцінка ринкової вартості об'єкту нерухомості для продажу.

Варіант 2. Оцінка ринкової вартості об'єкту нерухомості для здачі в оренду.

Варіант 3. Оцінка ринкової вартості свого комп'ютера.

Варіант 4. Оцінка ринкової вартості свого телефону (планшета).

Варіант 5. Оцінка ринкової вартості засобу пересування (машини, мотоцикла, велосипеда).

Варіант 6. Оцінка ринкової вартості об'єкту нерухомості для продажу.

Варіант 7. Оцінка ринкової вартості об'єкту нерухомості для здачі в оренду.

Варіант 8. Оцінка ринкової вартості свого комп'ютера.

Варіант 9. Оцінка ринкової вартості свого телефону (планшета).

Варіант 10. Оцінка ринкової вартості засобу пересування пересування (машини, мотоцикла, велосипеда).

Варіант 11. Оцінка ринкової вартості об'єкту нерухомості для продажу.

Варіант 12. Оцінка ринкової вартості об'єкту нерухомості для здачі в оренду.

Варіант 13. Оцінка ринкової вартості свого комп'ютера.

Варіант 14. Оцінка ринкової вартості свого телефону (планшета).

Варіант 15. Оцінка ринкової вартості засобу пересування (машини, мотоцикла, велосипеда).