Міністерство освіти і науки України

Черкаський державний технологічний університет

Кафедра програмного забезпечення автоматизованих систем

**ЗВІТ**

з лабораторної роботи №4

з предмету «Аналіз великих даних»

|  |  |
| --- | --- |
| Перевірив:  к. т. н., д. каф. ПЗАС  Рідкокаша А.А.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2019 р. | Виконав:  студент 1-го курсу  групи МПЗ-1904  Гаврилюк В. Є. |

Черкаси 2019

**Лабораторна робота №4**

**Тема:** Знайомство з Weka API для регресійного аналізу.

**Мета:** Ознайомитися та отримати навички практичного застосування методів бібліотеки Weka для регресійної класифікації даних.

**Завдання:** Розробити програму для побудови регресійної моделі даних.

|  |  |
| --- | --- |
| **Варіант** | **Завдання** |
| 5 | SMOreg |

**Короткі теоретичні відомості**

Weka (Waikato Environment for Knowledge Analysis) — вільне програмне забезпечення для аналізу даних та машинного навчання, написане на Java в університеті Ваїкато (Нова Зеландія), розповсюджується за ліцензією GNU GPL.

Weka — це набір засобів візуалізації та алгоритмів для аналізу даних і вирішення задач прогнозування, разом з графічною оболонкою для доступу до них.

Weka дозволяє виконувати такі завдання аналізу даних, як підготовку даних (preprocessing), відбір ознак (feature selection), кластеризацію, класифікацію, регресійний аналіз та візуалізацію результатів.

Основним інтерфейсом користувача є Explorer, хоча ті ж функціональні можливості підтримуються з командного рядка та інтерфейсу Knowledge Flow. Для систематичного порівняння різних алгоритмів машинного навчання використовується інтерфейс Experimenter. Він дозволяє порівнювати результати не лише різних алгоритмів на одному наборі даних, а й одного алгоритму на різних наборах даних.

Інтерфейс Explorer містить наступні панелі:

1. Панель попереднього опрацювання уможливлює імпорт даних з бази даних, текстових файлів у форматі CSV, а також попереднє опрацювання цих даних за допомогою різноманітних алгоритмів (фільтрів). Ці фільтри використовуються для трансформування даних, а також для видалення певних атрибутів.
2. Панель класифікації надає можливість застосувати алгоритми класифікації та регресійного аналізу до обраного набору даних, візуалізувати та оцінити результати, відобразити ROC криві тощо.
3. Панель асоціації надає доступ до методів, які дозволяють оцінити взаємозв'язки між атрибутами.
4. Панель кластеризації містить різноманітні методи кластеризації, наприклад метод кластеризації методом k-середніх, EM-алгоритм тощо.
5. Панель вибору атрибутів дозволяє ідентифікувати атрибути, які найбільш впливають на якість прогнозування.
6. Панель візуалізації відображає точкові діаграми.

**Хід роботи**

Для даної лабораторної роботи було створено файл diabetes.arff із наступними атрибутами:

% preg: Number of times pregnant

% plas: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

% pres: Diastolic blood pressure (mm Hg)

% skin: Triceps skin fold thickness (mm)

% insu: 2-Hour serum insulin (mu U/ml)

% mass: Body mass index (weight in kg/(height in m)^2)

% pedi: Diabetes pedigree function

% age: Age (years)

% class: Class variable (0 or 1)

@relation pima\_diabetes

@attribute 'preg' real

@attribute 'plas' real

@attribute 'pres' real

@attribute 'skin' real

@attribute 'insu' real

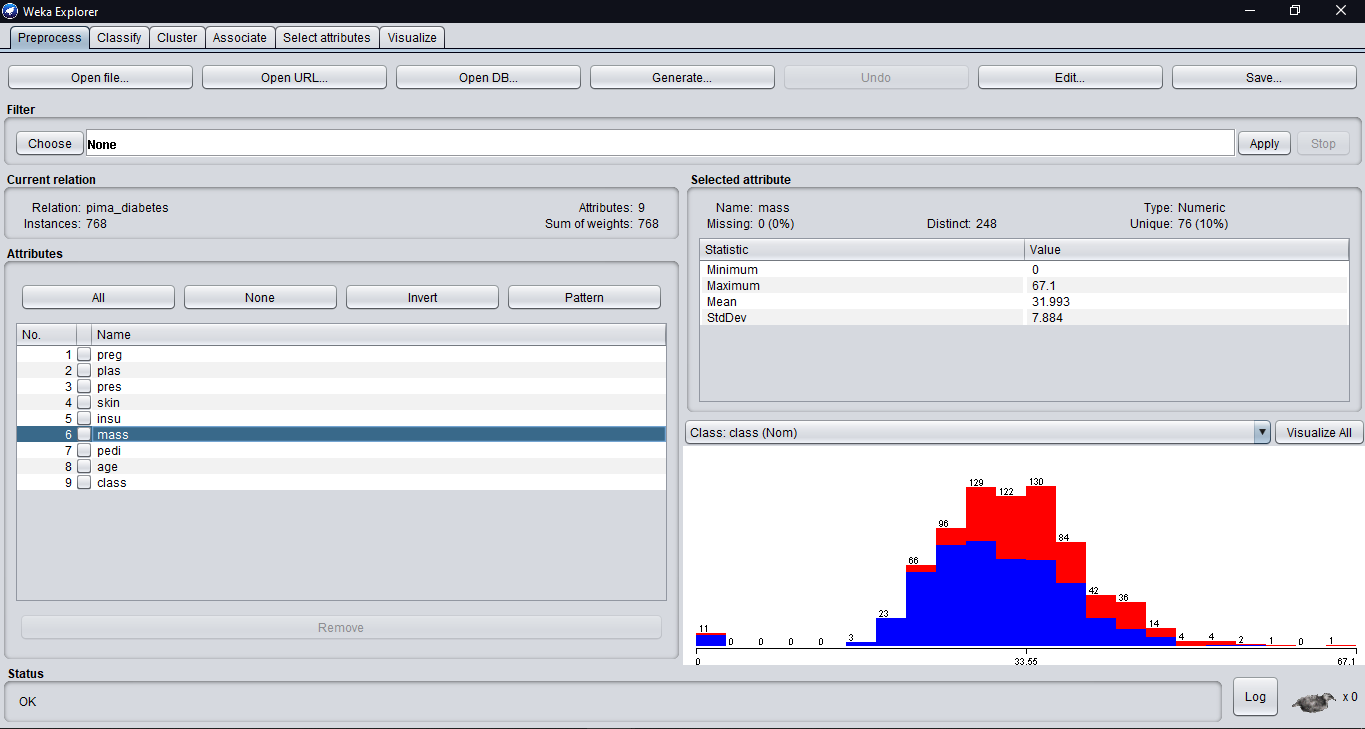
@attribute 'mass' real

@attribute 'pedi' real

@attribute 'age' real

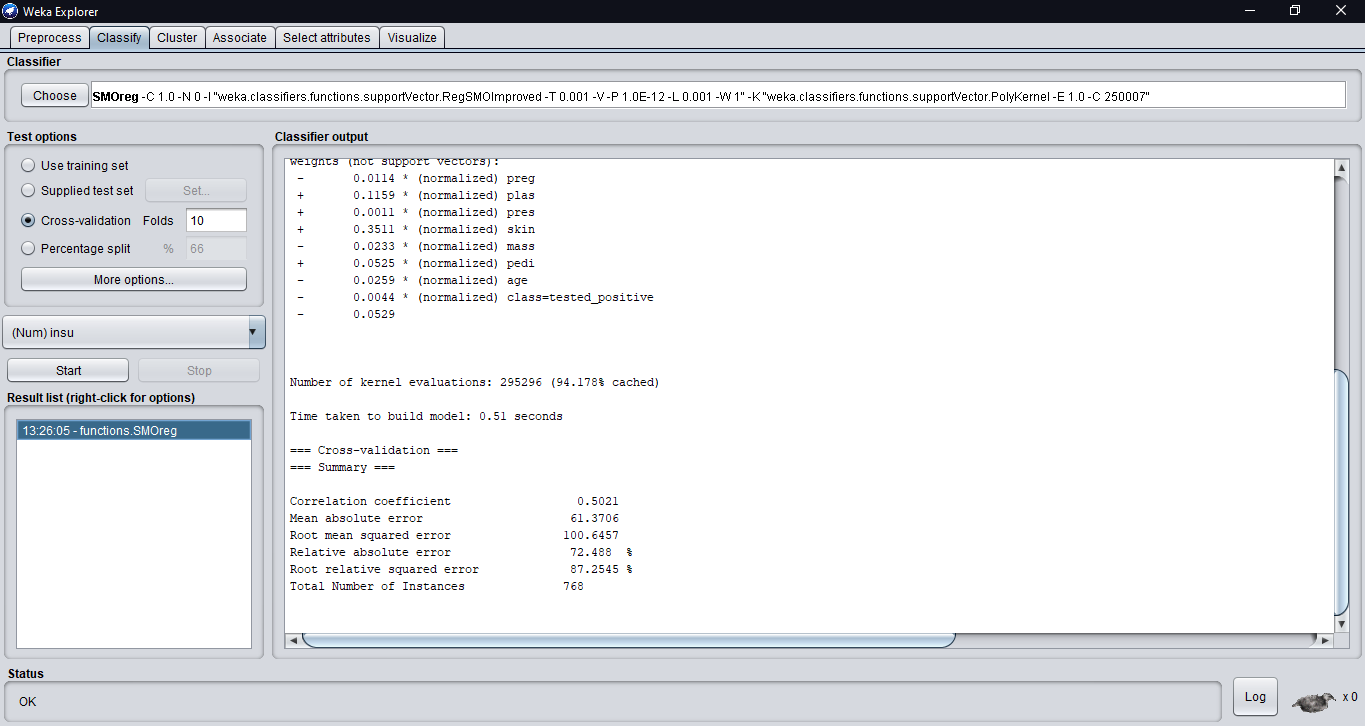
@attribute 'class' { tested\_negative, tested\_positive}

Після відкриття створеного файлу diabetes.arff в Weka Explorer можна побачити короткі відомості про дані, занесені у .arff-файл та атрибути створеної моделі. Також можно проглянути інформацію за певними атрибутами створеної моделі (рис. 1).



*Рис. 1 – Інформація про параметр моделі 'mass', що відображає ІМТ*

Для прогнозування рівня інсуліу у вагітних жінок під зі спадковим діабетом за результатами значень їх медичних обстежень було використано алгоритм SMOreg. Нижче зображено результат обрахунків та налаштування конфігурації для правильних розрахунків (рис. 2).



*Рис. 2 – Результат побудови SMOreg*

**Висновок:** Під час даної лабораторної роботи я ознайомився та отримав навички практичного застосування методів бібліотеки Weka для регресійної класифікації даних. Розробив програму для побудови регресійної моделі даних, яка спрогнозувала рівень інсуліну у вагітних жінок зі спадковим діабетом.