**Міністерство освіти і науки України**

**Національний Технічний Університет України**

**«Київський політехнічний інститут ім. Ігоря Сікорського»**

**Теплоенергетичний факультет**

**Кафедра автоматизації проектування енергетичних процесів і систем**

**Лабораторна робота №2**

з дисципліни

«Прикладні методи аналізу даних»

Виконав:

аспірант групи ТВ-71ф

Ляшенко М.М.

**Київ - 2018**

### Завдання: побудова класифікатора з використанням Weka

1. **Встановити середовище Weka**

Середовище було встановлено звідси:

* Завантаження та інсталяція Java: <https://java.com/en/>
* Завантаження та інсталяція Weka: <https://www.cs.waikato.ac.nz/ml/weka/downloading.html>

1. **Вивчити документацію по роботі з Weka.**

Встановлена програма вже включає в себе керівництво PDF (WekaManual.pdf). Керівництво включає детальні скріншоти. Також, в Інтернеті є велика кількість додаткової документації:

* Сторінка Weka: <http://weka.wikispaces.com/>
* Pentaho Data Mining Community Documentation: http://wiki.pentaho.com/display/DATAMINING/Pentaho+Data+Mining+Co mmunity+Documentation
* Часті запитання:  http://weka.wikispaces.com/Frequently+Asked+Questions

1. **Вибрати один з тренувальних наборів данних**

Була вибрана Pima Indians Diabetes Database.

Дані взяті з:

(a) Original owners: National Institute of Diabetes and Digestive and

Kidney Diseases

(b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)

Research Center, RMI Group Leader

Applied Physics Laboratory

The Johns Hopkins University

Johns Hopkins Road

Laurel, MD 20707

(301) 953-6231

(c) Date received: 9 May 1990

Декілька обмежень були віднесені до вибору цих екземплярів з більшої бази даних. Зокрема, всі пацієнти тут – жінки більше 21 року індійської спадщини Піма. ADAP – адаптивне процедура навчання, яка генерує і виконує цифрові аналоги перцептроноподібних пристроїв.

1. **Провести повне дослідження цього набору у відповідності з документацією по Weka.**

Кількість об’єктів – 768

Кількість атрибутів – 8+ класів

Атрибути:

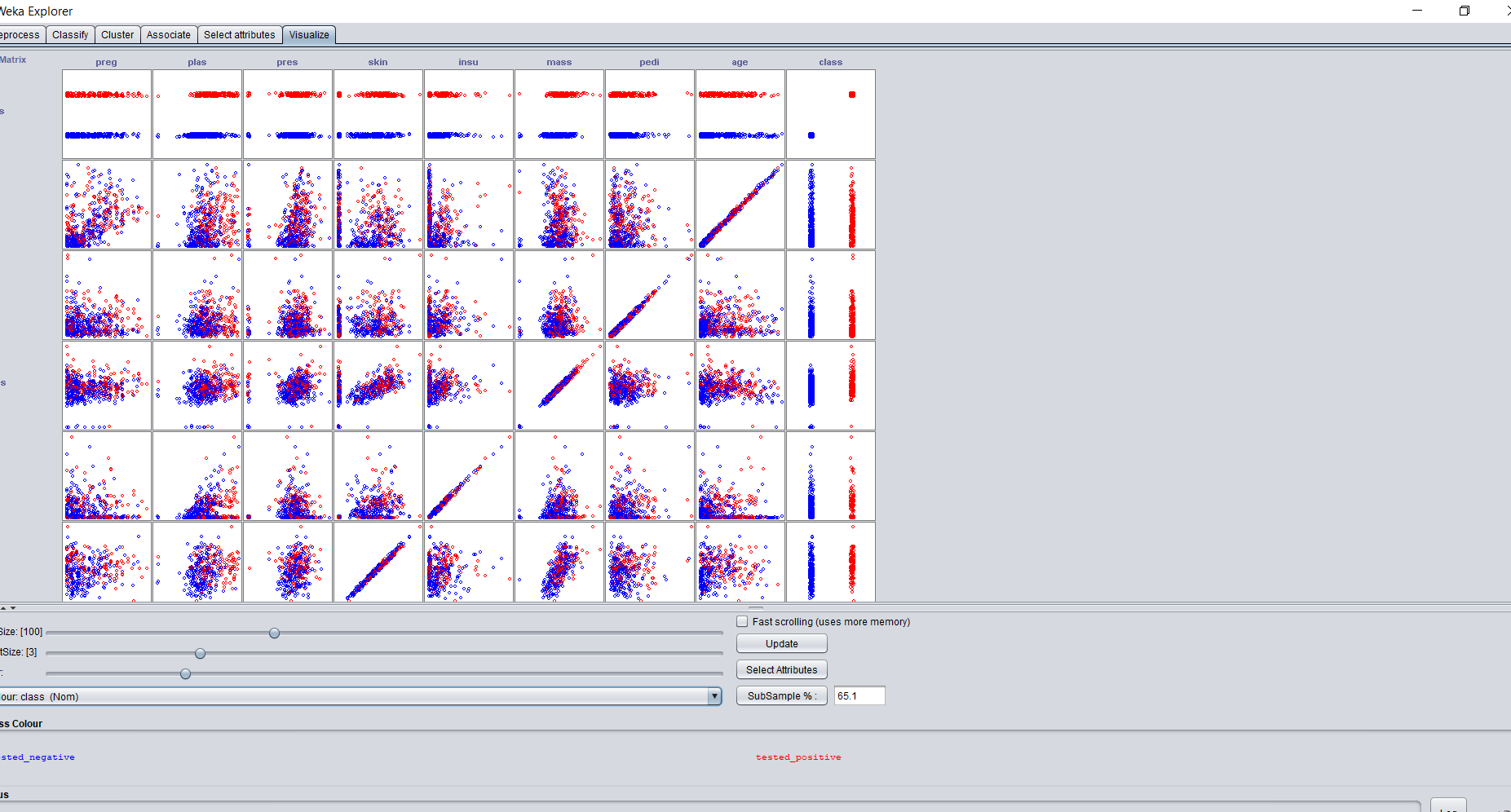
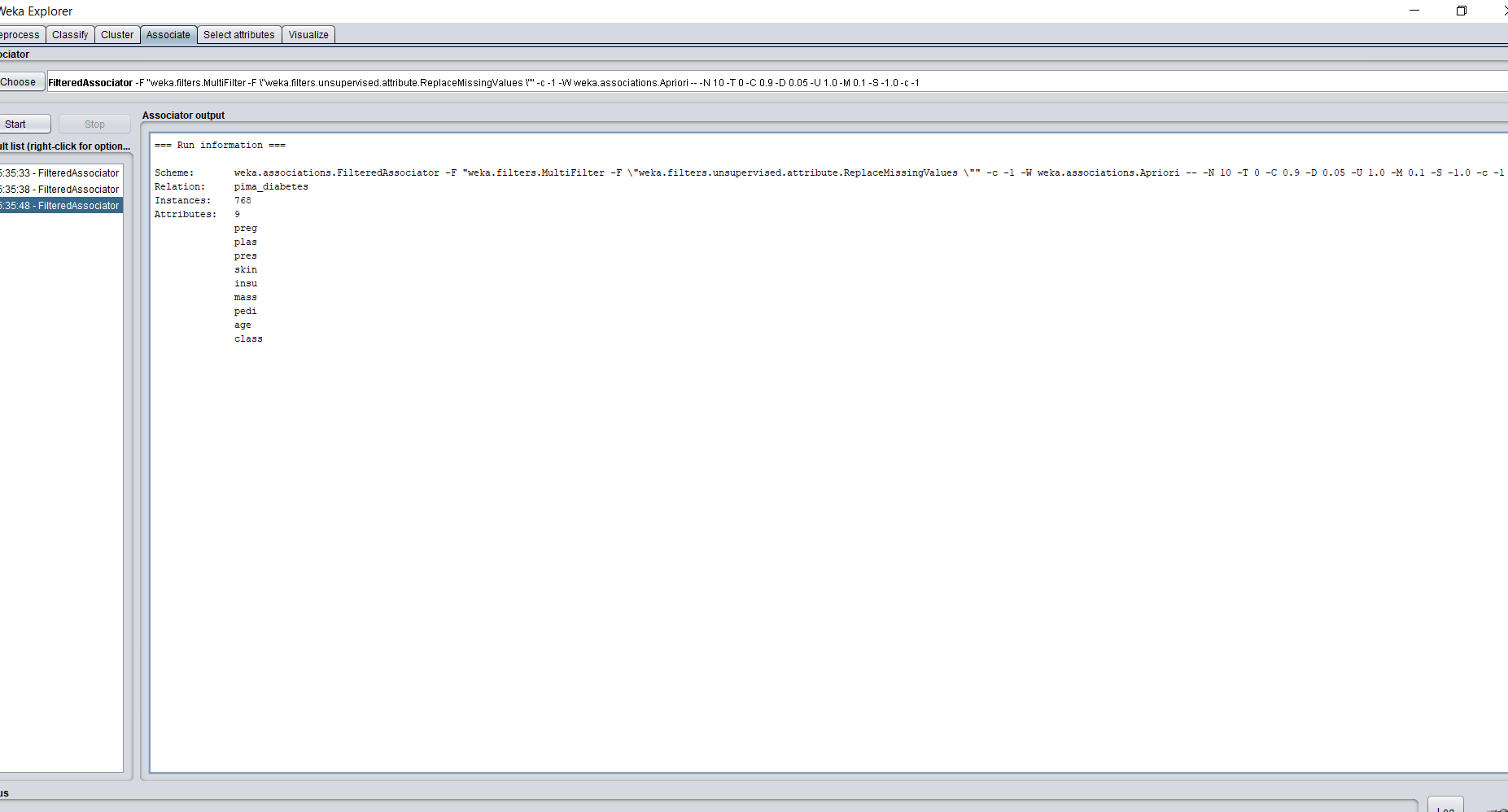
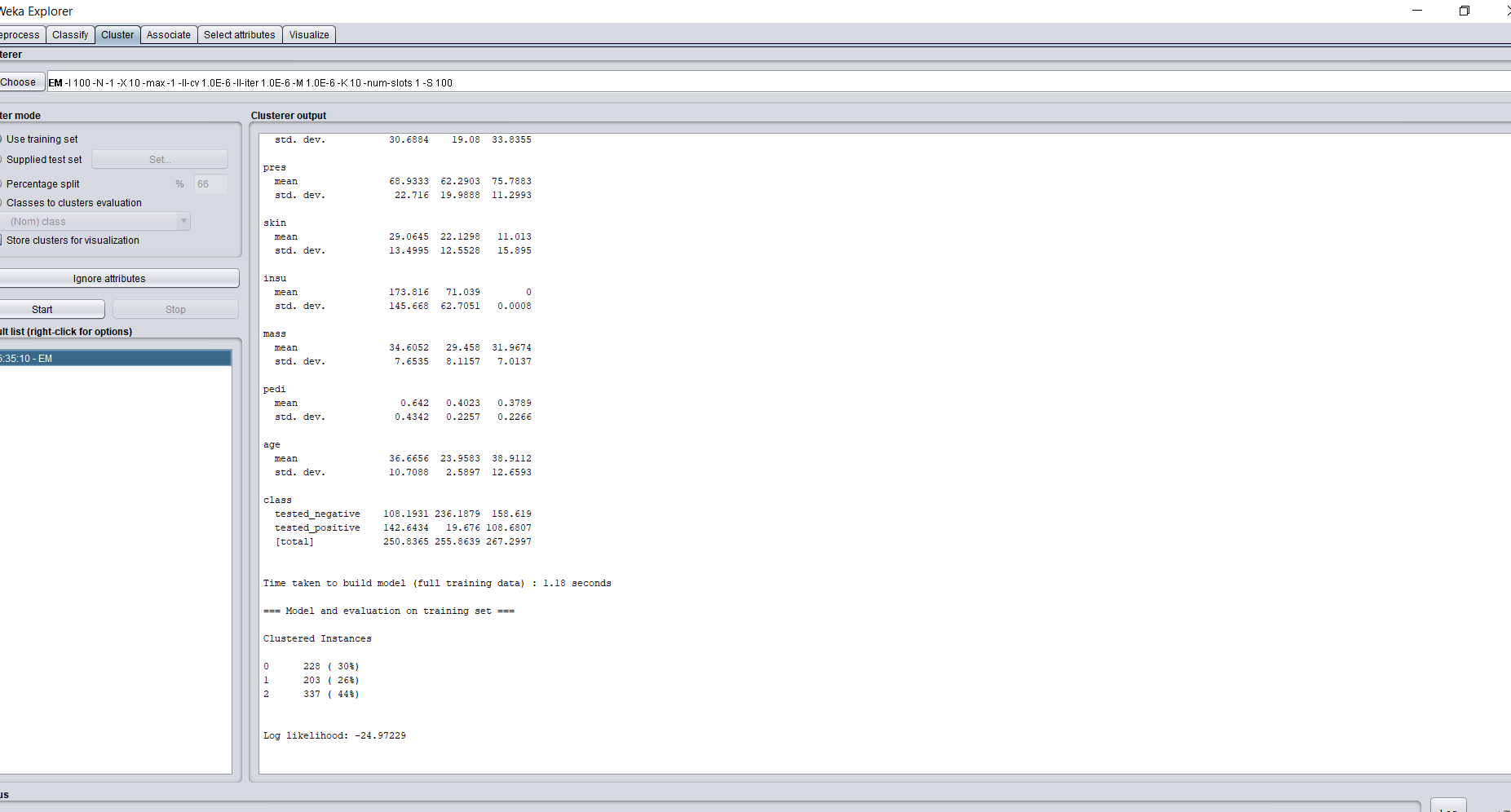
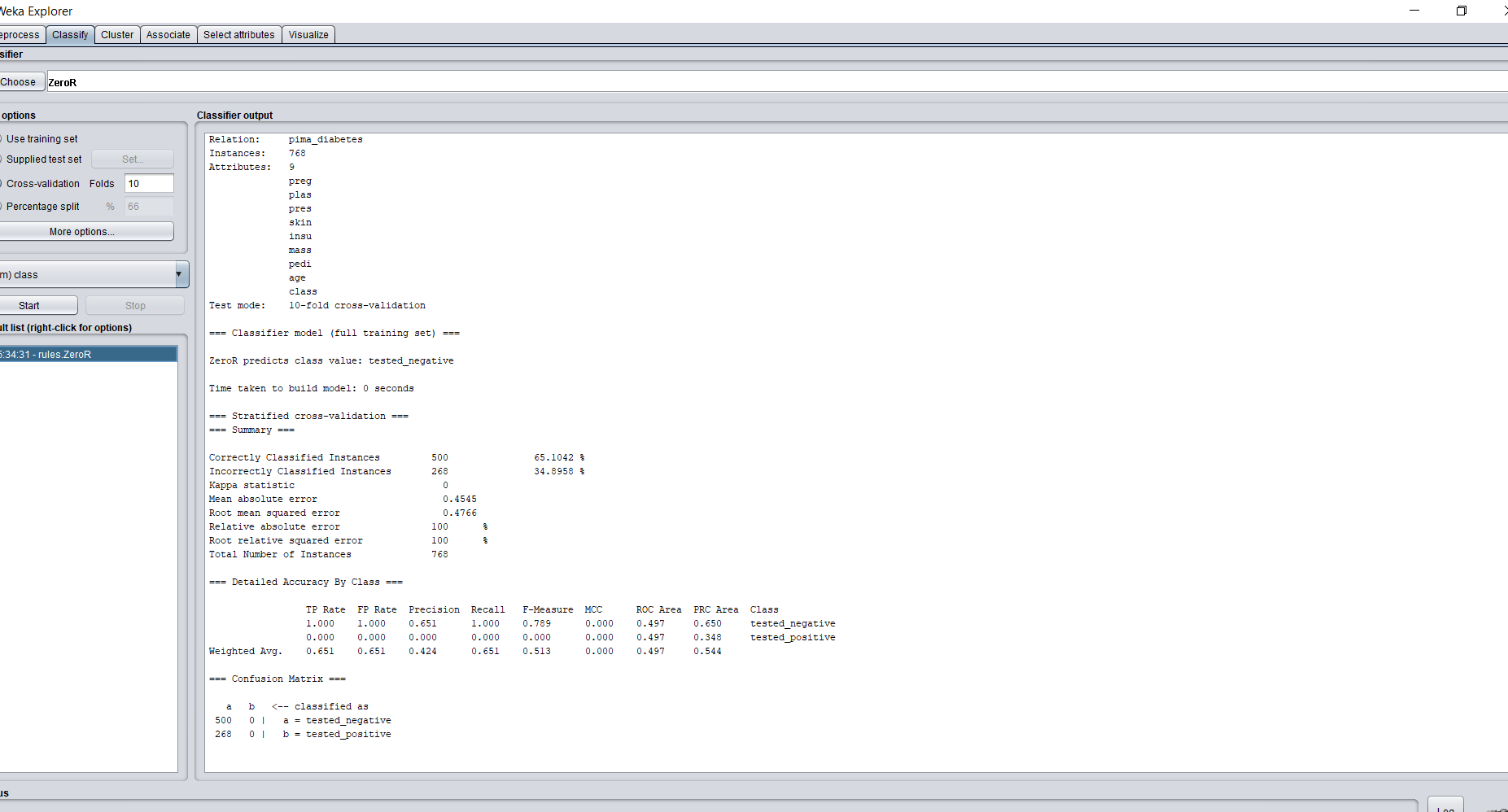
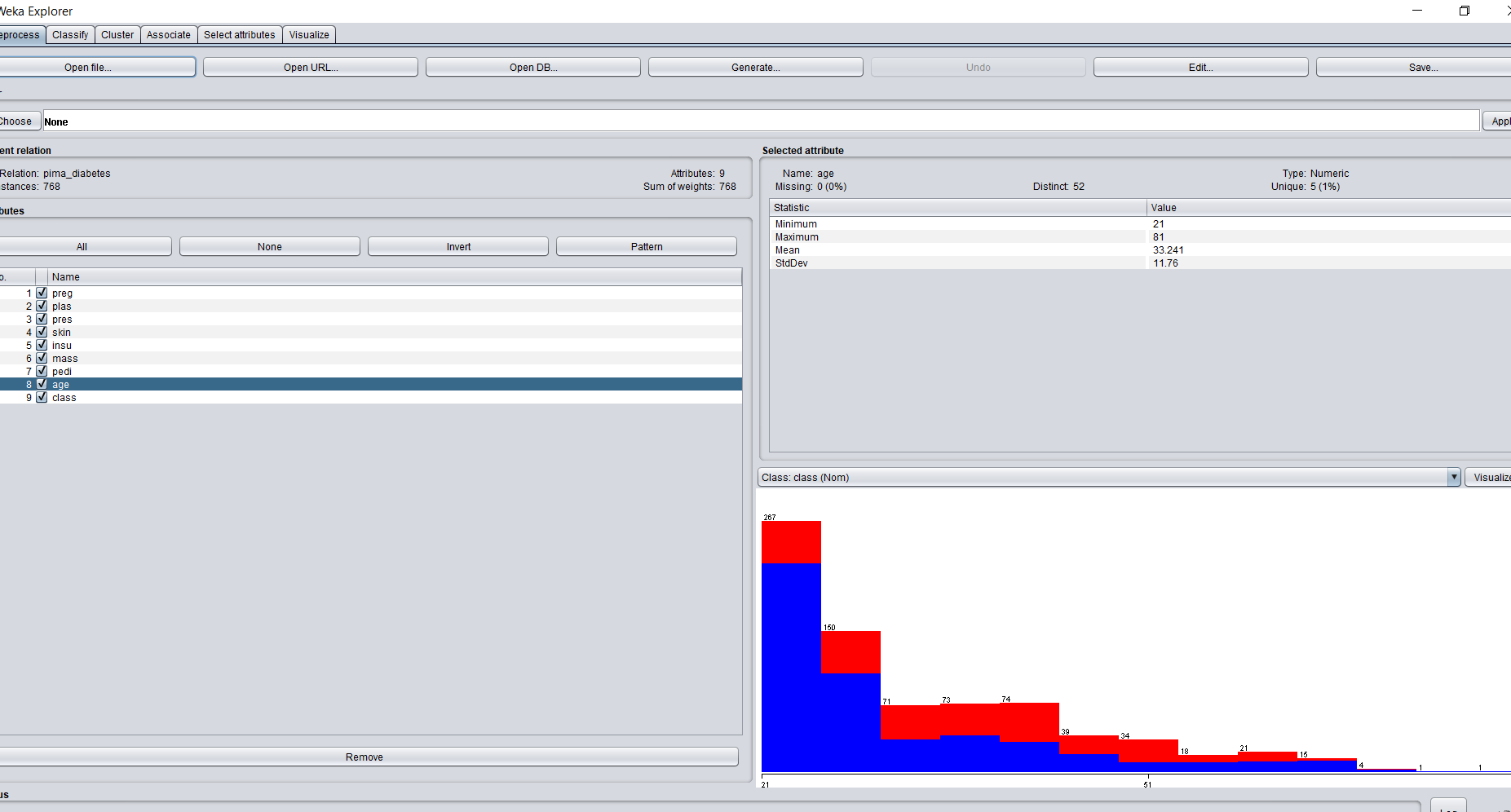
* Кількість вагітностей
* Концентрація глюкози в плазмі через 2 години у пероральному тесті на толерантність до глюкози
* Діастолічний кров'яний тиск (мм рт.ст.)
* Товщина складки шкіри трицепса (мм)
* 2-годинний сироватковий інсулін (мкО / мл)
* Індекс маси тіла (вага в кг / (висота в м) ^ 2)
* Функція родового діабету
* Вік (років)
* Класова змінна (0 або 1). Классова змінна 1 інтерпретується як «позитивна для діабету»

В резульатті:

0 – 500 об’єктів

1 – 268 об’єктів

Скріншоти аналізу:



### Висновки

Всі матераіли можна подивитися тут - <https://github.com/maxml/MachineLearning>

Були досліджені база даних діабету у індійців Піма (всього 768 записів). Майже у 125 дітей було від 3 до 5 дітей з данної вибірки, це другий результат, більше тільки очікувані 246 жінок з від 0 до 1 дитини. Майже у всіх досліджуваних жінок однаковий індекс тіла – 129-130. Також майже всі досліджувані жінки мали вік – 21-30 років (21-25 – 267, 25-30 - 150). На результатах кластеризації можна помітити, що кластеризації більш-менш усереднена по трьом кластерам, тільки кластер «2» при параметрі insu показує результат mean – 0, std. dev. 0.0008. Всі інші кластери по параметрам дають різницю в результатах не більше як в 2 рази. Саме це ми можемо побачити і на вкладці візуалізації – більшість результатів знаходиться по середині квадрата.