Міністерство освіти і науки, молоді та спорту України

Національний Технічний Університет України

«Київський Політехнічний Інститут»

Навчально-науковий комплекс

«Інститут прикладного системного аналізу»

Кафедра системного проектування

Лабораторна робота №3

з дисципліни «Інтелектуальний аналіз даних»

на тему:« **Знайомство з методами побудови дерев класифікації**»

Варіант №14

Виконав: студент 3 курсу

групи ДА-52

Плотніков С.O.

Перевірив: [Іщенко Г. В.](http://rozklad.kpi.ua/Schedules/ViewSchedule.aspx?v=327eda95-fa17-470c-a431-e67a04d790d4).

Київ 2018

***Мета роботи***: Ознайомитися та отримати навички побудови дерев класифікації за допомогою Data Mining GUI бібліотеки Weka.

*Метод классификации* (также известный как метод классификационных деревьев или деревьев принятия решений) - это алгоритм анализа данных, который определяет пошаговый способ принятия решения в зависимости от значений конкретных параметров. Дерево этого метода имеет следующий вид: каждый узел представляет собой точку принятия решения на основании входных параметров. В зависимости от конкретного значения параметра вы переходите к следующему узлу, от него – к следующему узлу, и так далее, пока не дойдете до листа, который и дает вам окончательное решение. Звучит довольно запутанно, но на самом деле метод достаточно прямолинеен. Давайте обратимся к конкретному примеру.

Листинг 1. Простое классификационное дерево

|  |  |
| --- | --- |
|  | [ Вы будете читать этот раздел?  ]            /              \          Да               Нет          /                 \  [Вы его поймете?]    [Не изучите]      /         \    Да          Нет    /             \  [Изучите]   [Не изучите] |

Это простое классификационное дерево определяет ответ на вопрос «Изучите ли вы принцип построения классификационного дерева?» В каждом узле вы отвечаете на соответствующий вопрос и переходите по соответствующей ветке к следующему узлу, до тех пор, пока не дойдете до листа с ответом «да» или «нет». Эта модель применима к любым сущностям, и вы сможете ответить, в состоянии ли эти сущности изучить классификационные деревья, с помощью двух простых вопросов. В этом и состоит основное преимущество классификационных деревьев – они не требует чрезмерного количества информации для создания достаточно точного и информативного дерева решений.

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning.

***Аналіз захворювання на діабет***

Інформація про дані та атрибути екземплярів даних.

Кількість атрибутів: 8 + клас

1. Кількість разів вагітності
2. Концентрація глюкози в плазмі при пероральному тестуванні проби на переносимість
3. Діастолічний артеріальний тиск
4. Товщина кожи трицепсу
5. 2-годинний сироватковий інсулін
6. Індекс маси тіла(кг/(м^2))
7. Родословна діабету(чи хворів хтось у сім’ї на діабет)
8. Вік(роки)
9. Клас: 0 або 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}

Кількість екземплярів даних: 768

Клас зі значенням 1 інтерпретується як tested positive for diabetes

|  |  |
| --- | --- |
| Значення класу | Кількість екземплярів |
| 0 | 500 |
| 1 | 268 |

@data

6,148,72,35,0,33.6,0.627,50,tested\_positive

1,85,66,29,0,26.6,0.351,31,tested\_negative

8,183,64,0,0,23.3,0.672,32,tested\_positive

1,89,66,23,94,28.1,0.167,21,tested\_negative

0,137,40,35,168,43.1,2.288,33,tested\_positive

5,116,74,0,0,25.6,0.201,30,tested\_negative

3,78,50,32,88,31,0.248,26,tested\_positive

10,115,0,0,0,35.3,0.134,29,tested\_negative

2,197,70,45,543,30.5,0.158,53,tested\_positive

8,125,96,0,0,0,0.232,54,tested\_positive

4,110,92,0,0,37.6,0.191,30,tested\_negative

10,168,74,0,0,38,0.537,34,tested\_positive

10,139,80,0,0,27.1,1.441,57,tested\_negative

1,189,60,23,846,30.1,0.398,59,tested\_positive

5,166,72,19,175,25.8,0.587,51,tested\_positive

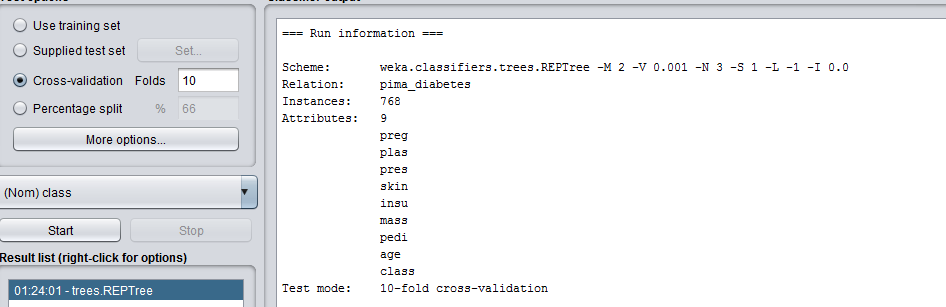
7,100,0,0,0,30,0.484,32,tested\_positive

0,118,84,47,230,45.8,0.551,31,tested\_positive

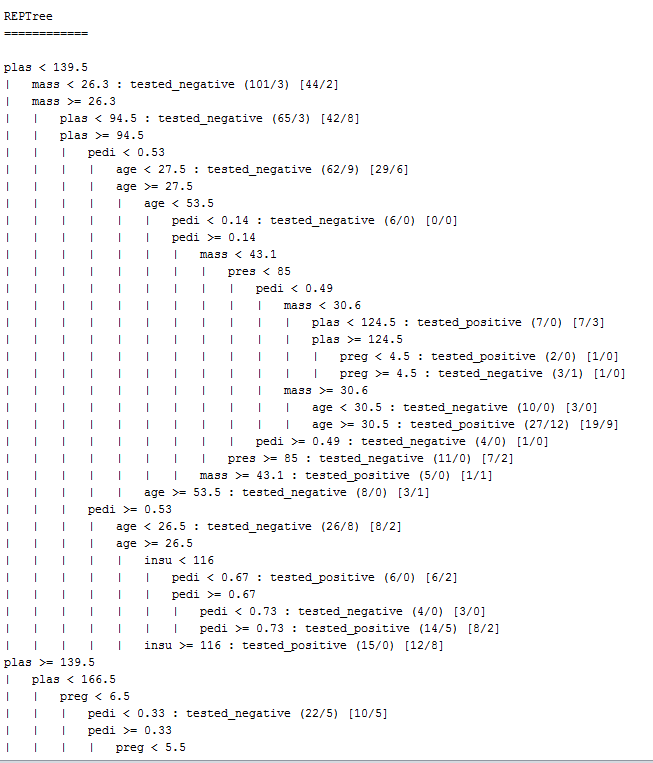
7,107,74,0,0,29.6,0.254,31,tested\_positive

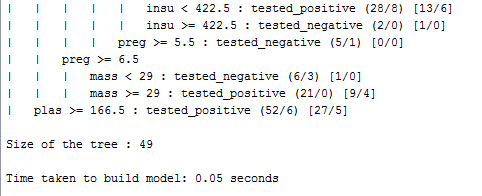
1,103,30,38,83,43.3,0.183,33,tested\_negative

………………………………………………………………………



**Побудова моделі класифікації у Weka Explorer**





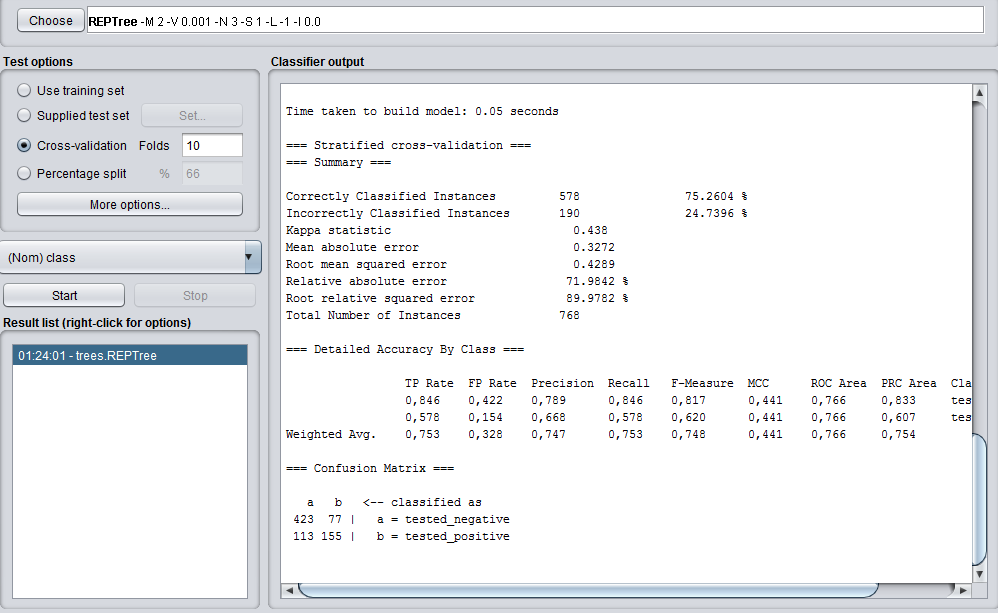


Таблица Confusion Matrix показывает количество ложноотрицательных

(77) и ложноположительных (113) распознаваний.

***Висновок:***

Зі значень, наведених у «Матриці достовірності» (Confusion Matrix), видно, що точність класифікації екземплярів з навчального набору дорівнює 75.2604% (423 екземплярів з 500 класу «tested negative» класифіковані коректно; 155 з 268 екземплярів класу «tested positive» класифіковані коректно).