Computer machine learning

COMP4388

Machine learning

Dr: Mustafa Jarrar

Second Project

- **Prepared by:**
  - Amir Eleyan-1191076.

- **Date:** Tuesday, May 24, 2022

Introduction

The goal of this project is to build a model using machine learning to predict “What is the predominant material used in the construction of the external walls of the dwelling” based on data collected under a survey called “Palestinian Expenditure and Consumption Survey” in 2011. About 30 columns were used as features (inputs to be used to predict the target value).

The machine learning algorithm used is the decision tree, especially C5.0, and this was done using the language R, which is a an open-source programming language that is widely used as a statistical software and data analysis tool.

Features

File name: Dwelling.dta

Input features:

Feature name	Description
H1	Type of housing unit
H2	Tenure of the housing unit
H5	How many rooms are there in dwelling
H6	How many sleeping rooms are used in dwelling
H8a	What is estimated rent value each month
H8b	Specify type of currency
H9a	Connection to Water
H9b	Connection to Electricity
H9c	Connection to Sewage system
H10	Availability of a kitchen
H11	Availability of a bathroom
H12	Availability of a toilet (WC):
H13_1	Main source of Cooking
H13_2	Main source of Heating
H13_3	Main source of Conditioner
H13_4	Main source of Oven
H13_5	Main source of Water heater
H14_1	Do several or all of house rooms and corridors, and kitchen suffer from the Dampness
H14_2	Do several or all of house rooms and corridors, and kitchen suffer from the Cold and difficult heating in winter

H14_3	Do several or all of house rooms and corridors, and kitchen suffer from the Poor ventilation
H14_4	Do several or all of house rooms and corridors, and kitchen suffer from the High heat in summer
H14_5	Do several or all of house rooms and corridors, and kitchen suffer from the Difficulty heating in winter
H18_1	Are family members in the dwelling or its surroundings exposed to Smoke, exhaust from cars
H18_2	Are family members in the dwelling or its surroundings exposed to Smoke, exhaust from industry
H18_3	Are family members in the dwelling or its surroundings exposed to Odors resulting from animals
H18_4	Are family members in the dwelling or its surroundings exposed to Odors resulting from sewage system water
H18_5	Are family members in the dwelling or its surroundings exposed to Odors resulting from
H18_6	Are family members in the dwelling or its surroundings exposed to General dust
H18_7	Are family members in the dwelling or its surroundings exposed to Dust or smells resulting from other sources
H18_8	Are family members in the dwelling or its surroundings exposed to Noise
H21_1	Availability of a car for the family
H21_5	Availability of a Cooking stove for the family
H21_6	Availability of a Dish washer for the family
H21_9	Availability of a Dehumidifier for the family
H21_16	Availability of a Computer for the family
H21_20	Availability of a Filter for the family

Output feature:

What is the main material used in building outside walls of housing unit	
1	Cleaned stone
2	Stone & cement
3	Old stone
4	Cement cob
5	Concrete
6	Mud
7	Other (specify)

Results:

```
Evaluation on training data (3453 cases):

Trial        Decision Tree
-----      ----------------
           Size      Errors

   0        148   708(20.5%)
   1         76   855(24.8%)
   2        103   891(25.8%)
   3        124   859(24.9%)
   4        122   931(27.0%)
   5        105   877(25.4%)
   6        120   958(27.7%)
   7        141   930(26.9%)
   8        149   843(24.4%)
   9        109   832(24.1%)
boost             553(16.0%)   <<
```

Figure 1: Model after 10 of boosting iterations

```
Attribute usage:

100.00% h1
100.00% h9a
100.00% h13_5
 99.86% h5
 98.99% h13_2
 93.89% h8b
 89.57% h21_6
 80.77% h13_4
 76.60% h8a
 71.10% h9c
 67.56% h21_1
 64.41% h9b
 63.94% h10
 62.73% h14_3
 62.18% h18_5
 59.80% h2
 59.19% h14_4
 57.52% h13_3
 56.21% h18_4
```

Figure 2: Percentage of each attribute was used, part-1

```
 54.21% h18_1
 51.17% h18_7
 50.97% h14_1
 49.23% h6
 47.32% h14_5
 44.05% h18_8
 43.96% h18_6
 38.52% h14_2
 36.55% h18_3
 33.85% h13_1
 33.16% h21_9
 27.14% h21_16
 12.86% h12
```

Figure 3: Percentage of each attribute was used, part-2

```
 (a)   (b)   (c)   (d)   (e)   (f)   (g)    <-classified as
----  ----  ----  ----  ----  ----  ----
 414     2     2   140                       (a): class Cleaned Stone
  32    61     3   117     1                 (b): class Stone & Cement
  20     3    58   110                   1   (c): class Old Stone
  29     3     6  2295                   2   (d): class Cement Cob
   1     1     2    63    57                 (e): class Concrete
                    11                       (f): class Mud
                     4                  15   (g): class Other
```

Figure 4: Confusion matrix

```
> predictions <- predict(model, x_testingDataSet)
> #calculate accuracy
> temp = predictions == y_testingDataSet
> accuracy = length(which(temp)) / length(temp) * 100.0
> sprintf("The accuracy= %.2f",accuracy)
[1] "The accuracy= 76.39"
```

Figure 5: Accuracy of the model in the testing data