Original Dataset:

Absenteeism at work

Abstract:

The database was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.

Source:

Original owners and donors: Andrea Martiniano (1), Ricardo Pinto Ferreira (2), and Renato Jose Sassi (3).

E-mail address:

andrea.martiniano@gmail.com (1) - PhD student;

log.kasparov@gmail.com (2) - PhD student;

sassi@uni9.pro.br (3) - Prof. Doctor.

Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management.

Number of instances: 740

Number of attributes: 21

Attribute Information:

1. Individual identification (ID)
2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioral disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

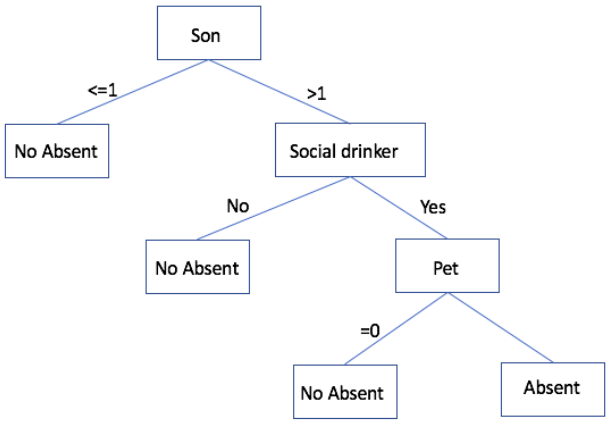
XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

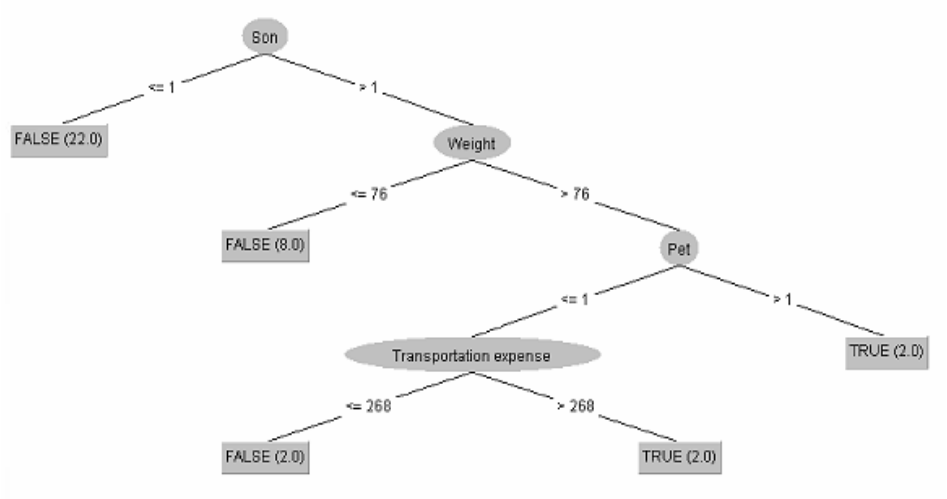
XVII Congenital malformations, deformations and chromosomal abnormalities
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX Injury, poisoning and certain other consequences of external causes
XX External causes of morbidity and mortality
XXI Factors influencing health status and contact with health services.
And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense
7. Distance from Residence to Work (kilometers)
8. Service time
9. Age
10. Work load Average/day
11. Hit target
12. Disciplinary failure (yes=1; no=0)
13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son (number of children)
15. Social drinker (yes=1; no=0)
16. Social smoker (yes=1; no=0)
17. Pet (number of pet)
18. Weight
19. Height
20. Body mass index
21. Absenteeism time in hours

Goal:
To see who might absent from work
My dataset:
I choose 10 attributes include ‘Transportation expense’, ‘Distance from Residence to Work’, ‘Age’, ‘Education’, ‘Son’, ‘Social drinker’, ‘Social smoker’, ‘Pet’, ‘Weight’, ‘Absenteeism time in hours’ and one label ‘Absent’ which is decided by my rule.
My rule:
A worker who has son >1 and is a social drinker and has pet >0 then he will absent from work

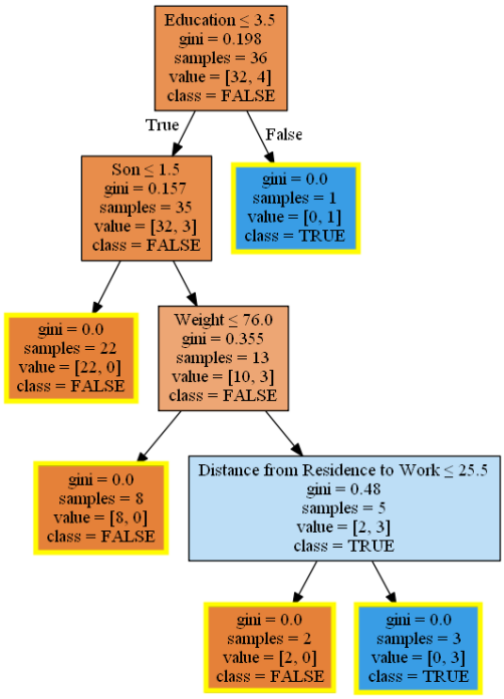
My decision tree:



The decision tree from WEKA J48:



The tree from python decision tree package:



Discussion:

Based on my rules four people will be considered as absent from work and WEKA shows the same answer as mine. However, WEKA uses two different attributes in the tree.

I think that is because as a human I uses my common sense to set the rule. A person that has many kids might take days off often, a person who drinks often might be absent from work because of a hangover or a person has many pets needs more time to accompany his pets. However, for J48 used by WEKA, it calculates every attribute based on the label. Then the gain is calculated that would result from the attribute. Then the best attribute that present selection criterion is found.

Python uses a different technique of GINI, which uses only one same attributes as myself and WEKA. The root of python's tree is education and it splits the tree by education ≤ 3.5. In my data, the attribute education means high school = 1, graduate = 2, postgraduate = 3, master and doctor = 4. Even when python splits the result right, the number 3.5 doesn't mean anything. In data mining research, it is hard to let the computer to know all the attributes, even if we get the right answer. Sometimes it is hard for us to check whether each layer in the tree makes a rational determination. I believe this shows that human users can select attributes that very closely match an optimized computer solution, at least in some cases.