

International Journal of Computer Science and Mobile Computing



A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 2, February 2017, pg.8 – 16

Body Mass Index (BMI) Data Analysis and Classification

Mahesh Mardolkar

Principal, Bharatesh College of Computer Applications, Belagavi, Karnataka, India- 590016

principal@bharateshbca.com

Abstract: *The body mass(weight) and the height of an individual are the attributes used to derive Body Mass Index(BMI) or Quetelet Index. The body mass divided by square of the body height gives BMI and expressed in unites kg/m^2 Ross Quinlan developed C4.5 algorithm to generate a decision tree, extension of Quinlan's earlier ID3 algorithm is C4.5, C4.5 is always referred as statistical classifier as it generates decision tree which can be used for classification.*

A BMI application is used to illustrate C5.0 to construct diagnostics rules, each case in a BMI application belongs to classes (Underweight, Normal, Overweight, Obese), there are 12 attributes in this application, but C5.0 can handle any number of attributes.

Keywords: *Body Mass Index (BMI), C4.5, C5.0*

Introduction

Adolphe Quetelet a Belgian astronomer, sociologist, mathematician and statistician devised BMI which he called as social physics, due to its simplicity BMI is widely used in preliminary diagnosis. Person's thinness or thickness can be measured using BMI, which allows to discuss weight related problems with patients. The current BMI recommendation values are, BMI above $30 kg/m^2$ suggest person is obese, a value between 25 to 30 indicates the person is overweight, a value between 18.5 to 25 indicates optimal weight and below 18.5 suggest the person is underweight, if the dimensions of the body doubles which increases the mass as BMI is proportional to mass and inversely proportional to square of height so BMI also doubles, many

reports of BMI is seen in taller people, the WHO regards a BMI of less than 18.5 as underweight and may indicate malnutrition, an eating disorder.

In children BMI is calculated in same way as for adults, but compared with typical values of children with same age.

C4.5 Algorithm

Ross Quinlan developed C4.5 algorithm to generate a decision tree, extension of Quinlan's earlier ID3 algorithm is C4.5, C4.5 is always referred as statistical classifier as it generates decision tree which can be used for classification.

The concept of information entropy is used to build a tree from a set of training data, classified samples are made available in training data, which consists of p-dimensional vector.

C4.5 splits the set of samples into subset which is normalized information gain, to make the decision we choose the attribute with highest normalized information gain. C4.5 algorithm is again implemented on smaller subsets.

Pseudocode –

1. Base cases is checked.
2. Find the normalized information gain received from splitting on a, for each attribute.
3. The highest normalized information gain be a best.
4. To split a best create decision node.
5. Child nodes are created by splitting a_best, and recurs on sublist.

Improvement over ID3 algorithm-

- ➔ Both continuous and discrete attributes can be handled.
- ➔ Missing attributes in training data can be handled.
- ➔ Attributes with different costs can also be handled.
- ➔ Tree can be pruned after creation which removes branches which are not helpful.

Improvements in C5.0 algorithm-

Quinlan created C5.0 which has number of improvements over C4.5 and these are

- ➔ Speed – with several orders of magnitude C5.0 is much faster than C4.5
- ➔ Memory – memory efficiency is seen more in C5.0
- ➔ Smaller – smaller decision trees are found in results of C5.0
- ➔ Boosting – boosting improves the accuracy.
- ➔ Weighting – different cases are weighted
- ➔ Winnowing – unhelpful attributes are removed automatically by C5.0.

BMI(Body Mass Index) –

The body mass(weight) and the height of an individual are the attributes used to derive Body Mass Index(BMI) or Quetelet Index. The body mass divided by square of the body height gives BMI and expressed in unites kg/m^2 , mass is in kilograms and height is in meters. BMI categorizes the person as underweight, normal, overweight or obese. The different categories of BMI is mentioned in the table1.

A BMI application is used to illustrate c5.0 to construct diagnostics rules, each case in a BMI application belongs to classes(Underweight, Normal, Overweight, Obese), there are 12 attributes in this application, but C5.0 can handle any number of attributes.

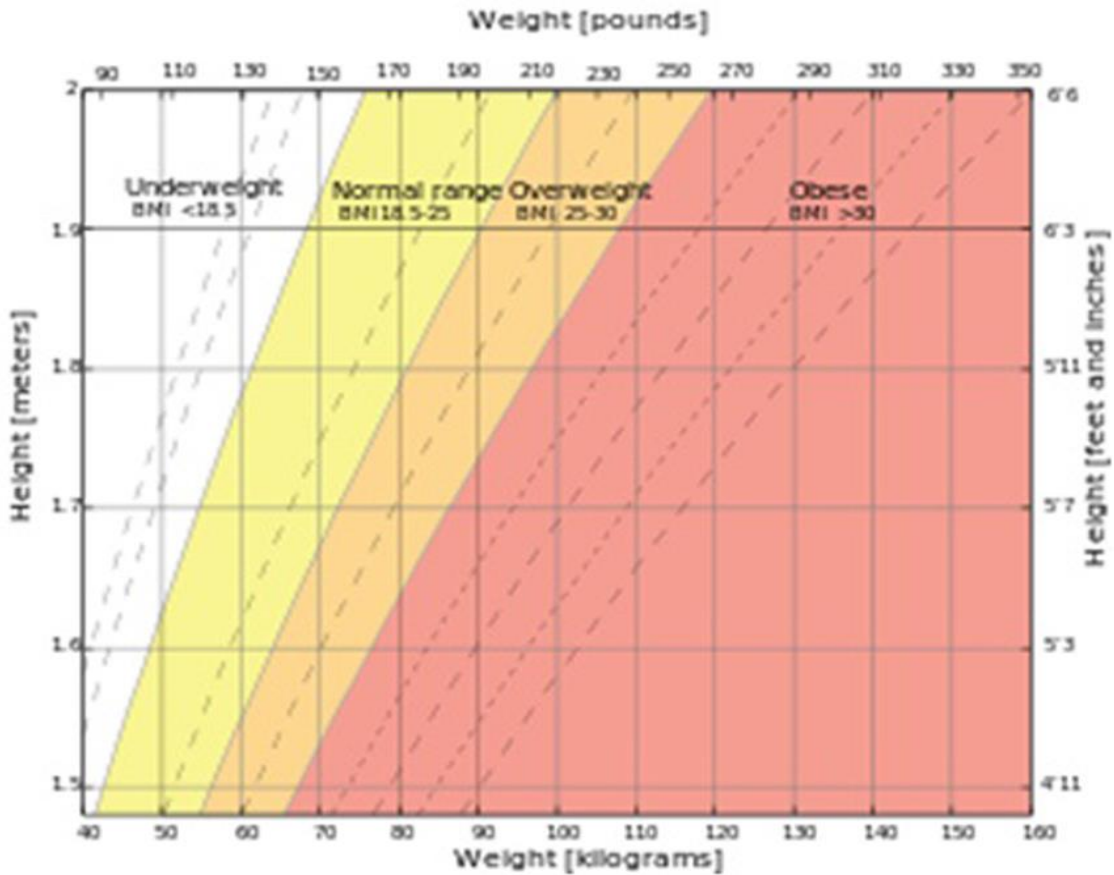


Fig1 – Graphical classification of BMI

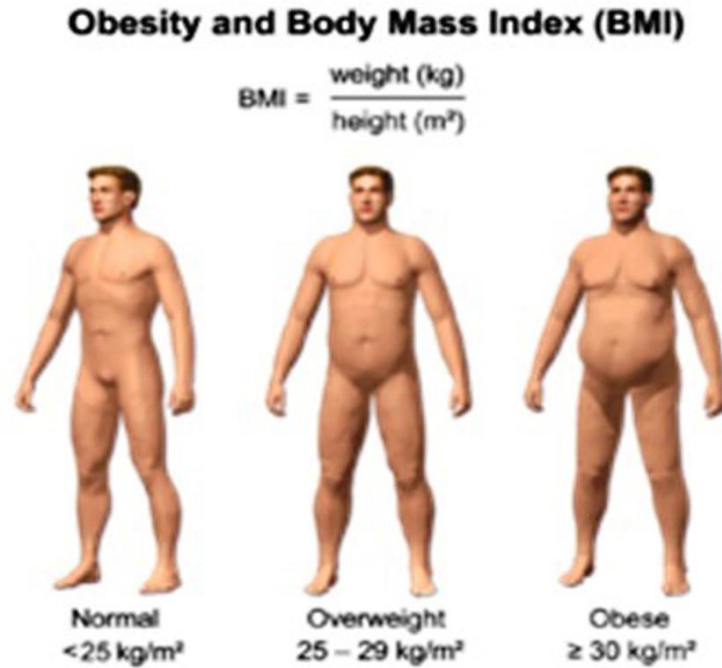


Fig2 – Normal, Overweight and Obese categories of BMI

Gory	BMI (kg/m ²)		BMI Prime	
	from	to	from	to
Very severely underweight		15.0		0.60
Severely underweight	15	16	0.60	0.64
Underweight	16	18.5	0.64	0.74
Normal (healthy weight)	18.5	25	0.74	1.0
Overweight	25	30	1.0	1.2
Obese Class I (Moderately obese)	30	35	1.2	1.4
Obese Class II (Severely obese)	35	40	1.4	1.6
Obese Class III (Very severely obese)	40		1.6	

Table 1 – Categories of BMI

Application files –

Filestem is the short name given to every C5.0 application, we will use BMI filestem for this illustration. The applications have name of the form filestem. extension, extension identifies the content of the file, below is the C5.0 extension summary.

Name	Description of the application's attribute	Required
Data	Cases used to generate a classifier	Required
Test	Unseen cases used to test a classifier	Optional
Cases	Cases to be classified subsequently	Optional
Costs	Differential misclassification costs	Optional
Tree	Decision tree classifier produced by c5.0	Output
Rules	Ruleset classifier	Output

Name File-

The name file is the first file that describes the attributes and classes. The BMI.name looks like this

Diagnosis | the target attribute

Case No: label

Name: label

Age: continuous

Sex: m, f

Height: continuous

Weight: continuous

BMI: continuous

Underweight: u, f

Normal: n, f

Overweight: ow, f

Obese: ob, f

Test: +, -

Data File-

The training cases is provided by second essential file BMI. data which consists of one or more lines, values are separated by commas and anything after '|' is ignored, any value missing or unknown is represented by '?' and not applicable values are denoted by 'N/A'. Cases added in BMI. data

Attributes	Case1	Case2	Case3
Case No				
Name				
Age				
Sex				
Height				
Weight				
BMI				
Underweight				
Normal				
Overweight				
Obese				
Test				

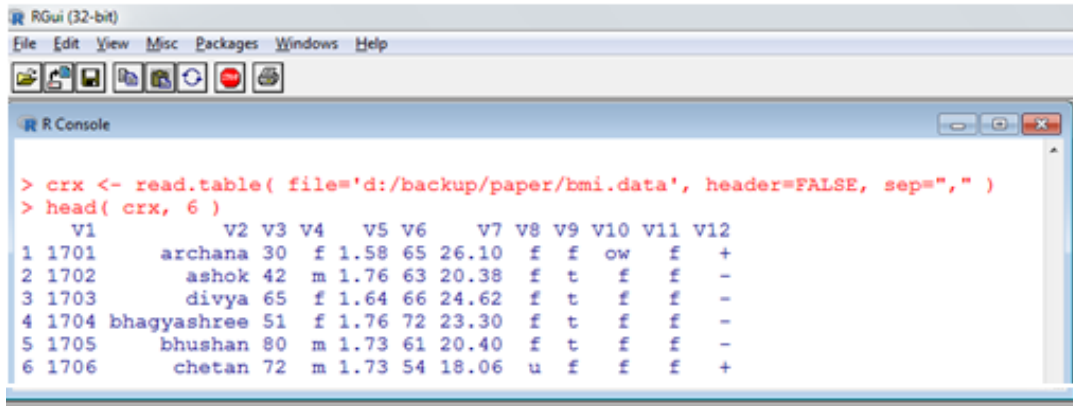
Implementing C50 package in R programming

The BMI data is taken as an example for creating decision tree using C50 package, in the given data set there are 30 records which is accessed directly from a file as

```
crx <- read.table( file=url, header=FALSE, sep="," )
```

the head() function is used to see first few lines of the file

```
head( crx, 6 )
```



```

> crx <- read.table( file='d:/backup/paper/bmi.data', header=FALSE, sep="," )
> head( crx, 6 )
  V1      V2 V3 V4   V5 V6   V7 V8 V9 V10 V11 V12
1 1701  archana 30 f 1.58 65 26.10 f f ow  f  +
2 1702   ashok 42 m 1.76 63 20.38 f t  f  f  -
3 1703   divya 65 f 1.64 66 24.62 f t  f  f  -
4 1704 bhagyashree 51 f 1.76 72 23.30 f t  f  f  -
5 1705   bhushan 80 m 1.73 61 20.40 f t  f  f  -
6 1706   chetan 72 m 1.73 54 18.06 u f  f  f  +
    
```

To randomize the rows we need to shuffle the rows and split the data to get the feature matrix X and target vector y.

```

crx <- crx[ sample( nrow( crx ) ), ]
X <- crx[,1:11]
y <- crx[,12]
    
```

we can use the training set out of given data for X and y as

```

trainX <- X[1:30,]
trainy <- y[1:30]
    
```

C50 package in R

The process of installing C50 package and accessing it through library() function is given below.

```

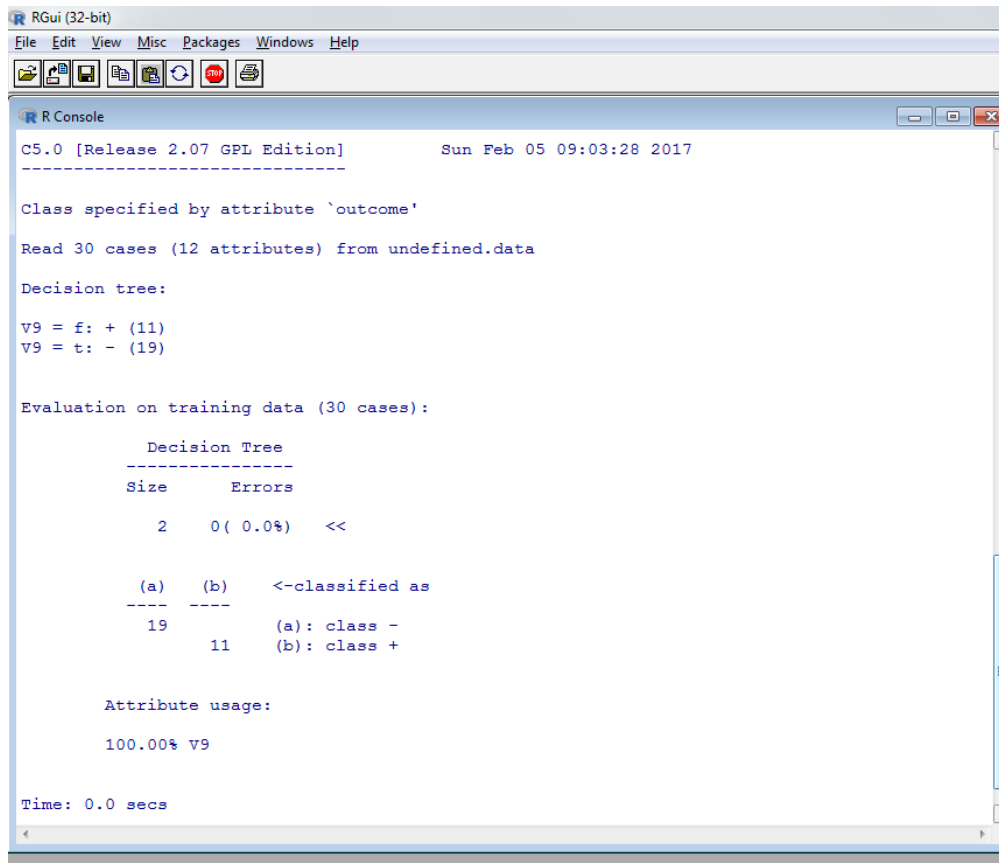
install.packages("C50")
library(C50)
    
```

to build a model and later find the summary of output as

```

model <- C50::C5.0( trainX, trainy )
summary( model )
    
```

the tree splits the given data of V12 variable of + and – values.



```
RGui (32-bit)
File Edit View Misc Packages Windows Help
R Console
C5.0 [Release 2.07 GPL Edition] Sun Feb 05 09:03:28 2017
-----
Class specified by attribute `outcome'
Read 30 cases (12 attributes) from undefined.data
Decision tree:
V9 = f: + (11)
V9 = t: - (19)
Evaluation on training data (30 cases):
      Decision Tree
      -----
      Size      Errors
      2      0( 0.0%)  <<
      (a)  (b)  <-classified as
      ---  ---
      19      11  (a): class -
                   (b): class +
Attribute usage:
100.00% V9
Time: 0.0 secs
```

Conclusion:

Too much of body fat means obesity, Obesity related health risk are

- > High Blood Pressure
- > Heart Disease and Stroke
- > Cancer
- > Diabetes
- > Gallstones
- > Gout
- > Osteoarthritis
- > Breathing problem

The body also includes weight from muscle, body water and bones, obesity also means the weight of a person is greater than normal for his or her height, Body Mass Index (BMI) is used to diagnose a person to be obese. The C50 package of R programming analyses the data and classifies them to understand the level of obesity among the cases diagnosed.

References

- [1] Daniel T. Larose & Chantal D. Larose, Discovering knowledge in Data, Second Edition John Wiley
- [2] Trever Hstie, Robert Tibshirani, & Jerome Friedman (2001). The Elements of Statistical Learning, First Edition Springer
- [3] Learning R – Richard Cotton – O’ REILLY
- [4] R – Peter Dalgaard Introductory Statistics with– Second Edition Springer
- [5] https://en.wikipedia.org/wiki/Body_mass_index
- [6] <http://connor-johnson.com/2014/08/29/decision-trees-in-r-using-the-c50-package/>
- [7] <https://medlineplus.gov/obesity.html>