DMAT Lab 04

2.5% of the module (if selected)

Student Name:Guy Erne Assika

Student Number:

In this worksheet you will achieve the following outcomes

1. Load data into Weka (csv)
2. Fix any issue that may arise.
3. Understand the data.
4. Preprocess the data.
5. Use Weka for classification

# Submissions

You will submit an edited version of this document complete with

* appropriately edited filename
* appropriate added student name and number at the top
* appropriate screen shots to illustrate that work has been completed with a single sentence of explanation below each screen short.

# Tasks

## Download the data titanic.csv

## Load the data into Weka.

**You will face some errors, identify the reason and fix it.**

(remember, real world data is not as clean as you expect)

## Understanding your data.

1. How many instances are there?

1308

1. How many attributes?

14

1. What are the types of the attributes?

Numeric, Nominal, String,

1. Give brief description of each of the attributes, can you make a quick decision which are useful in classifying the surviving class?

pclass Passenger Class = NUMERIC – 3 DISTINCT 0 MISSING

(1 = 1st; 2 = 2nd; 3 = 3rd)

survival Survival = NUMERIC – 2 DISTINCT 0 MISSING

(0 = No; 1 = Yes)

name Name NOMINAL- 1306 DISTINCT .. not useful in classifying...each are unique

sex Sex = 2 DISTINCT 0 MISSING

age Age = NUMERIC 98 DISTINCT , 263 MISSING

sibsp Number of Siblings/Spouses Aboard NUMERIC 7 DISTINCT and 0 MISSING

parch Number of Parents/Children Aboard = NUMERIC 8 DISTINCT , 0 MISSING

ticket Ticket Number = String, Not useful as ticket number is unique

fare Passenger Fare, NUMERIC 1 missing 281 distinct.. replace missing value with mean

cabin Cabin = 1013 missing so not useful to replace NOMINAL Attribute

embarked Port of Embarkation = NOMINAL AND 2 MISSING.

(C = Cherbourg; Q = Queenstown; S = Southampton)

boat Lifeboat STRING TYPE, only survived has boat attributes

body Body Identification Number (not useful as only dead bodies have it)

home.dest Home/Destination(NOMINAL)

## Pre-process the data.

1. Do all the pre-processing you think necessary to predict the class (survived). Document your work.

Fixed csv file by fixing quotation marks in name and destination address so that it can be imported.

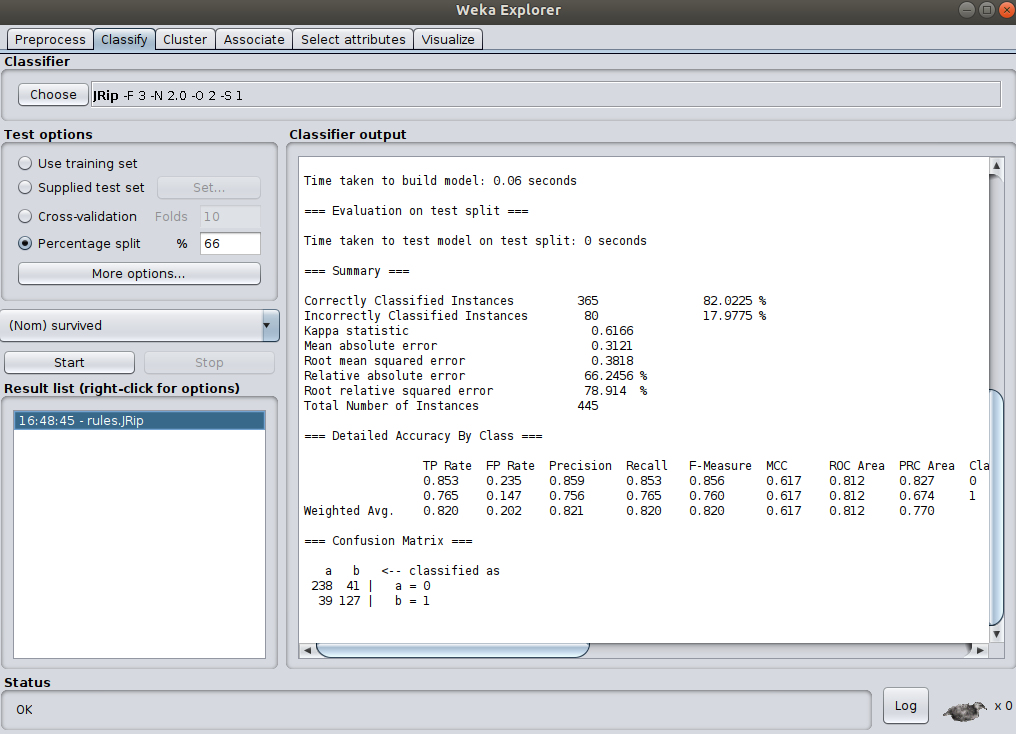
Dropped Name, Ticket, Fare, Cabin, homedest from dataset as it is not important.

Replaced missing values

Changed survived to nominal from numeric

## Classification using Weka.

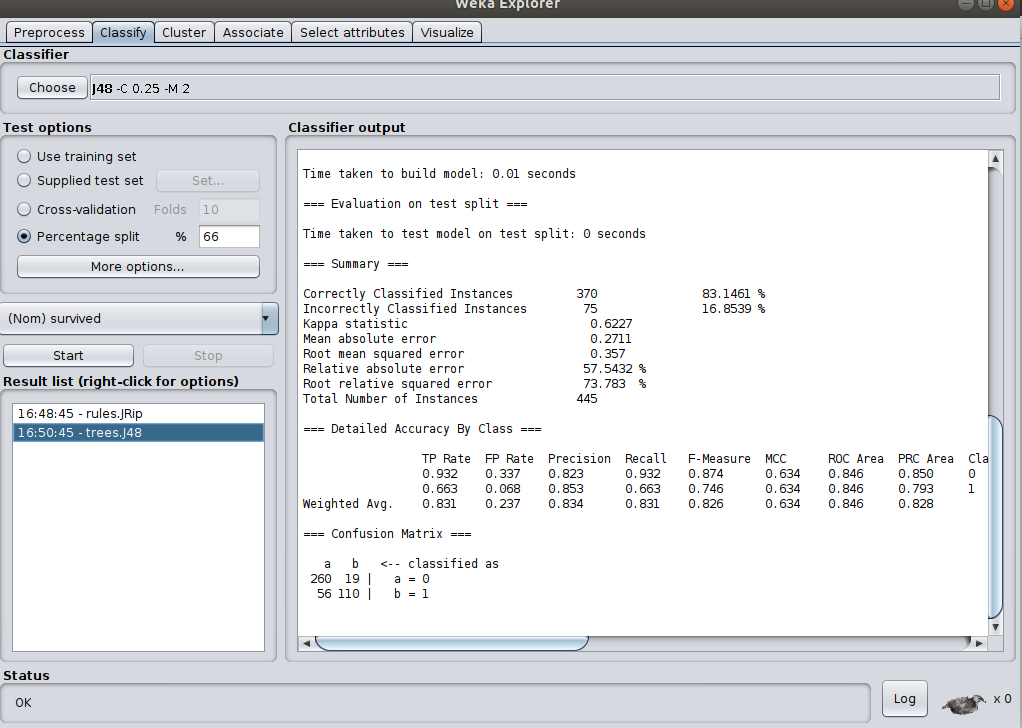
1. Use a classifier or two as your benchmark. Justify your choice.



Used Jrip classifier gave 82% accuracy.

1. Use J48 classifier. Vary the hyperparameters and report your findings.

83% accuracy achieved with j48 classifier



Submit on moodle within the time specified.

Data Description

pclass Passenger Class

(1 = 1st; 2 = 2nd; 3 = 3rd)

survival Survival

(0 = No; 1 = Yes)

name Name

sex Sex

age Age

sibsp Number of Siblings/Spouses Aboard

parch Number of Parents/Children Aboard

ticket Ticket Number

fare Passenger Fare

cabin Cabin

embarked Port of Embarkation

(C = Cherbourg; Q = Queenstown; S = Southampton)

boat Lifeboat

body Body Identification Number

home.dest Home/Destination