Homework Assignment 2 - Rule Induction/Decision Tree

1. This is an **individual assignment**. Submit a **MS WORD** or **pdf** with your name.
2. For questions 2 and 3, DO NOT put all your screen shots into an Appendix. The Screen shots should be in line in your report, so that the reader can readily look at the screenshots.
3. Make use of all the model evaluation techniques you know to draw conclusions.
4. **Inductive learning** is a process whereby the learner discovers rules by observing **examples**.

Q1: (Need NOT Submit. For your practice.): The following data set shows how different individuals have been prescribed contact lenses based on the available attribute data. The objective is to build a decision tree classifier using this data. (Data in **contact lens data.xlsx**)

1. How many classes are there? How many attributes?
2. Perform the necessary analysis to decide on which attribute should be chosen for split at the root of the tree. Note that you are asked to do this **just once at the root of the tree**. You are **NOT** asked to build the tree any further. Excel is a good tool to carry out your analysis. Show your work and justify your answer. When there are more than two classes, Entropy can be more than one. With three classes maximum entropy is 1.5849
3. Submit the Excel file in which you made the calculations and summarize the findings in the report.

Table 1.1 Contact Lens Data

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Q2 (40 pts): Please follow the Decision Tree/Rule Induction workshop document carefully in completing this assignment**. Inductive learning** is a process whereby the learner discovers rules by observing **examples**. The learning used in decision trees is Inductive Learning.

1. Complete the **SPSS workshop on Decision Trees**. This workshop is on BB. I have enclosed the document to this assignment as well. This workshop uses the **Churn.txt** dataset provided to you.
2. Your report (maximum three pages), should include:
   - Problem statement.
   - Your brief observations from a Data Audit.
   - What was done and analysis based on screenshots.
   - Managerial conclusions free of technical jargon. How can management make use of the model?
   - Screenshots, highlighting key values.
     1. Model stream
     2. Output from the Analysis note. Include the coincidence matrices, and model accuracy report.

Churn.txt Data description

In this workshop we use data from a telecommunications company, ***churn.txt***. The file contains records for 1477 of the company's customers who have at one time purchased a mobile plan. It includes such information as length of time spent on local, long distance and international calls, the type of billing scheme and a variety of basic demographics, such as age and gender. The customers fall into one of three groups: current customers, involuntary leavers, and voluntary leavers. We want to use data mining to understand what factors influence whether an individual remains as a customer or leaves for an alternative company. The data are typical of what is often referred to as a churn example (hence the file name).

churn.txt contains information from a telecommunications company. The data are comprised of customers who at some point have purchased a mobile phone. The primary interest of the company is to understand which customers will remain with the organization or leave for another company.

ChurnTrain.txt and **ChurnValidate.txt** were created by splitting the records from churn.txt into two files, one used during training and one for model validation.

The files contain the following fields:

ID	Customer reference number
LONGDIST	Time spent on long distance calls per month
International	Time spent on international calls per month
LOCAL	Time spent on local calls per month
DROPPED	Number of dropped calls
PAY_MTHD	Payment method of the monthly telephone bill
LocalBillType	Tariff for locally based calls
LongDistanceBillType	Tariff for long distance calls
AGE	Age
SEX	Gender
STATUS	Marital status
CHILDREN	Number of Children
Est_Income	Estimated income
Car_Owner	Car owner
CHURNED	(3 categories) Current – Still with company Vol – Leavers who the company wants to keep Invol – Leavers who the company doesn't want

Q3 (60pts): This question requires the file ***Risk.txt***, which is provided to you. A description of the data is enclosed.

Build a Decision Tree Classifier (C5.0) to predict the field ***Risk*.** Build your model and submit the model along with model evaluation in the form of a report. Use Windows snipping tool to copy and paste your stream, evaluation results etc.

Your report (max 3 pages) should include:

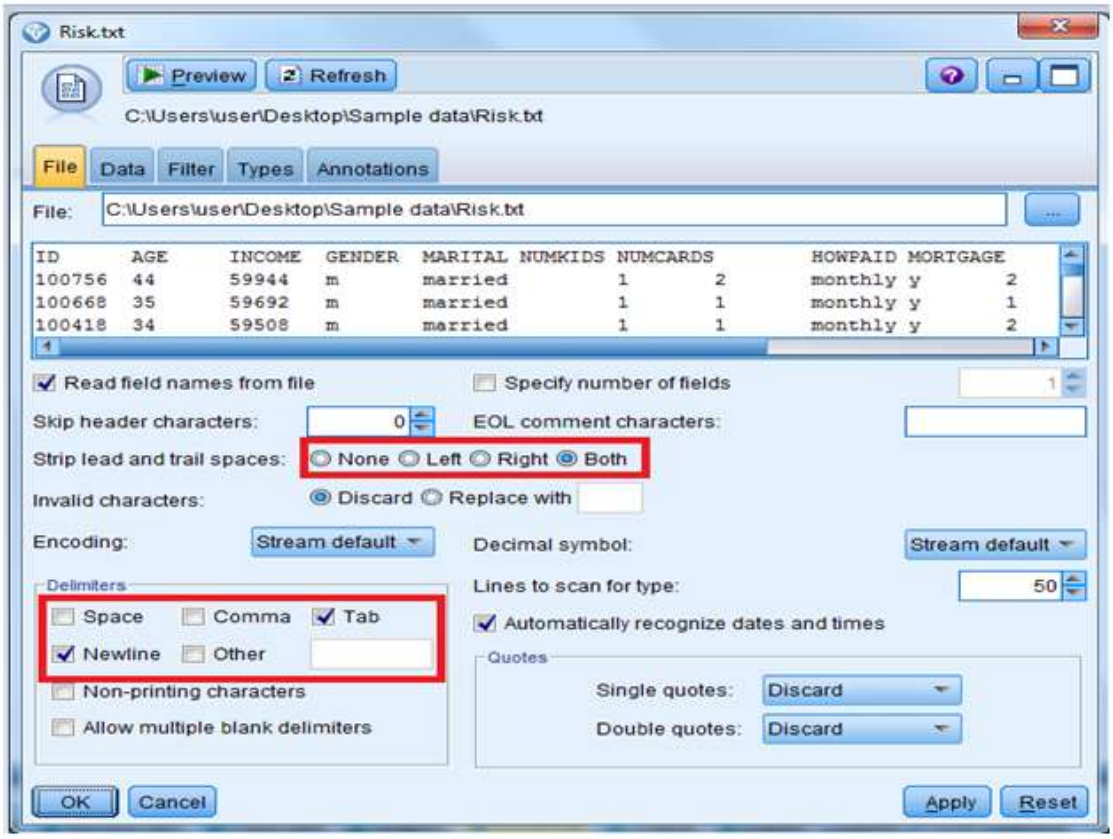
1. Problem statement. (What is the business problem being addressed?)
2. Brief summary of your observations from a Data Audit.
3. What was done and analysis based on the model evaluation..
4. Managerial conclusions free of jargon. How can management make use of the model?
5. Screenshots, highlighting key values. Include:
    a. Model stream.
    b. Model summary, predictor importance and expanded Decision tree.
    c. Performance Analysis.

Risk.txt contains information from a risk assessment study in which customers with credit cards were assigned to one of three categories: good risk, bad risk-profitable (some payments missed or other problems, but profitable for the issuing company), and bad risk-loss. In addition to the risk classification field, a number of demographic attributes are available for about 2,500 cases. We want to predict credit risk from the demographic fields. The file contains the following fields:

ID	ID number
AGE	Age
INCOME	Income in British pounds
GENDER	Gender
MARITAL	Marital status
NUMKIDS	Number of dependent children
NUMCARDS	Number of credit cards
HOWPAID	How often is customer paid by employer (weekly, monthly)
MORTGAGE	Does customer have a mortgage?
STORECAR	Number of store credit cards
LOANS	Number of outstanding loans
RISK	Credit risk category

4) Optional exercise: (need not submit): In SPSS modeler, C&RT is another decision tree algorithms. **C&RT** is like C5.0, but it uses Gini Index instead of Entropy to measure the homogeneity at a node. You can try and get familiar with the algorithm. All documentation is on black board. You will learn that C&RT can also be used for Regression by building a **Regression Tree.** I will introduce this later in this course.

How to read the Risk.txt file. Notice that the file is **not Comma separated**. It is **TAB separated file**. Follow the instructions below.



If you have difficulty reading the Risk.txt file, an Excel version of the file is available to you.