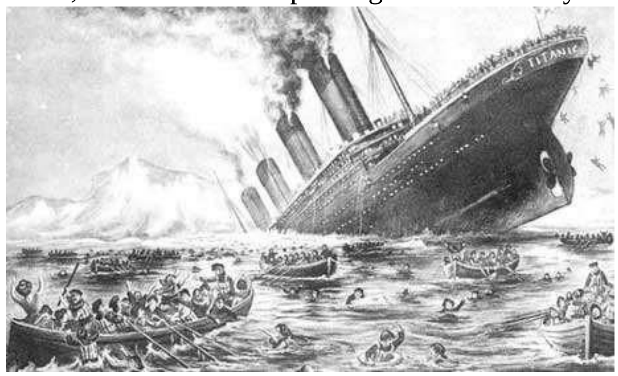
Assignment – Bayesian Classification

The **titanic dataset**[1] describes the survival status of individual passengers on the Titanic. The principal source for data about Titanic passengers is the *Encyclopedia Titanica*. The dataset used here were begun by a variety of researchers. One of the original sources is Eaton & Haas (1994) *Titanic: Triumph and Tragedy*, Patrick Stephens Ltd, which includes a passenger list created by many researchers and edited by Michael A. Findlay.



The variables on the dataset are all nominal: pclass, survived, age, and gender.

- pclass refers to passenger class (1st, 2nd, 3rd, crew), and is a proxy for socioeconomic class.
- age is dichotomized at adult vs. child.
- gender is male or female.
- survived is yes or no

The dataset has 2201 instances with no missing data (quite convenient). It is called titanic.xlsx

<u>PROBLEM 1</u>

I have randomly sampled 32 records out of the original dataset (called titanic_reduced.xlsx)

a) **The task is to <u>manually</u> build a Naive Bayes classifier that, by learning from previously collected data, is able to produce predictions on new demographic data. You must also smooth the model to deal with zero-frequency issues**

Following excel has been prepared as solution attached in "TitanincReduced.xlsx" file

pclass	age	gender	survived
1st	adult	male	no
2nd	adult	female	no
3rd	adult	male	no
3rd	adult	male	no
3rd	adult	male	no
3rd	adult	male	no
3rd	adult	male	no
3rd	adult	male	no
3rd	adult	male	no
3rd	adult	male	no
3rd	child	female	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
1st	adult	female	yes
1st	adult	female	yes
1st	adult	female	yes
1st	adult	female	yes
2nd	adult	female	yes
2nd	adult	female	yes
2nd	adult	male	yes
2nd	child	female	yes
2nd	child	female	yes
3rd	adult	male	yes
crew	adult	female	yes
crew	adult	female	yes
crew	adult	male	yes

Pclass probabilities	Row Labels	Count of pclass	Probability		
	no	**19**	**19/32**	**0.59**	p(Survived = no)
	1st	1	1/19	0.05	p(pclass = 1st | no)
	2nd	1	1/19	0.05	p(pclass = 2nd | no)
	3rd	9	9/19	0.47	p(pclass = 3rd | no)
	crew	8	8/19	0.42	p(pclass = crew | no)
	yes	**13**	**13/32**	**0.41**	p(survived = yes)
	1st	4	4/13	0.31	p(pclass = 1st | yes)
	2nd	5	5/13	0.38	p(pclass = 2nd | yes)
	3rd	1	1/13	0.08	p(pclass = 3rd | yes)
	crew	3	3/13	0.23	p(pclass = crew | yes)
	Grand Total	**32**			
Age probabilities	**Row Labels**	**Count of age**			
	no	**19**			
	adult	18	18/19	0.95	p(age = adult | no)
	child	1	1/19	0.05	p(age = child | no)
	yes	**13**			
	adult	11	11/13	0.85	p(age = adult | yes)
	child	2	2/13	0.15	p(age = child | yes)
	Grand Total	**32**			
Gender probabilities	**Row Labels**	**Count of gender**			
	no	**19**			
	female	2	2/19	0.11	p(gender = female | no)
	male	17	17/19	0.89	p(gender = male | no)
	yes	**13**			
	female	10	10/13	0.77	p(gender = female | yes)
	male	3	3/13	0.23	p(gender = male | yes)
	Grand Total	**32**			

Since there are no zero frequencies in the data, smoothing is not required.

[1] This is an all-nominal-features, no-missing-data dataset typically used in machine learning to assess the performance of classifiers.

b) Suppose that you look at a given individual record: (crew, adult, male). What would your Naïve Bayes model predict about the fate of this individual[2]?
Note1: as a recommendation, use Excel, it makes your life easier with calculations.
Note 2: do not partition the data in training and validation sets, use all the data for training (this is just a toy exercise for you to compute the probabilities, and we only have 33 records in this case).
Note 3: You don't have to be an experienced data miner to predict the fate of a male adult crew member ☺.

Let the record: crew, adult, male be X => X = (pclass=crew, age=adult, gender=male)

p(X | no)*p(no) = p(pclass=crew | no)*p(age=adult | no)*p(gender=male | no)*p(survived = no)

from the above excel figure (or attached file "titanicReduced.xlsx)

= 8/19 * 18/19 *17/19 *19/32 = 0.212

p(X | yes)*p(yes) = p(pclass=crew | yes)*p(age=adult | yes)*p(gender=male | yes)*p(survived = yes)

= 3/13 * 11/13 * 3/13 * 13/32 = 0.018

P(X | no) > p(X | yes) so the record is classified as not survived.

Computing probability of prediction
P(X | no) = 0.212/(0.018+0.212) = 0.920 = 92%

P(X | yes) = 0.018/(0.018+0.212) = 0.079 = 7.9%

Hence the record is classified as No with a confidence of 92% according to the Naïve Bayes model.

---

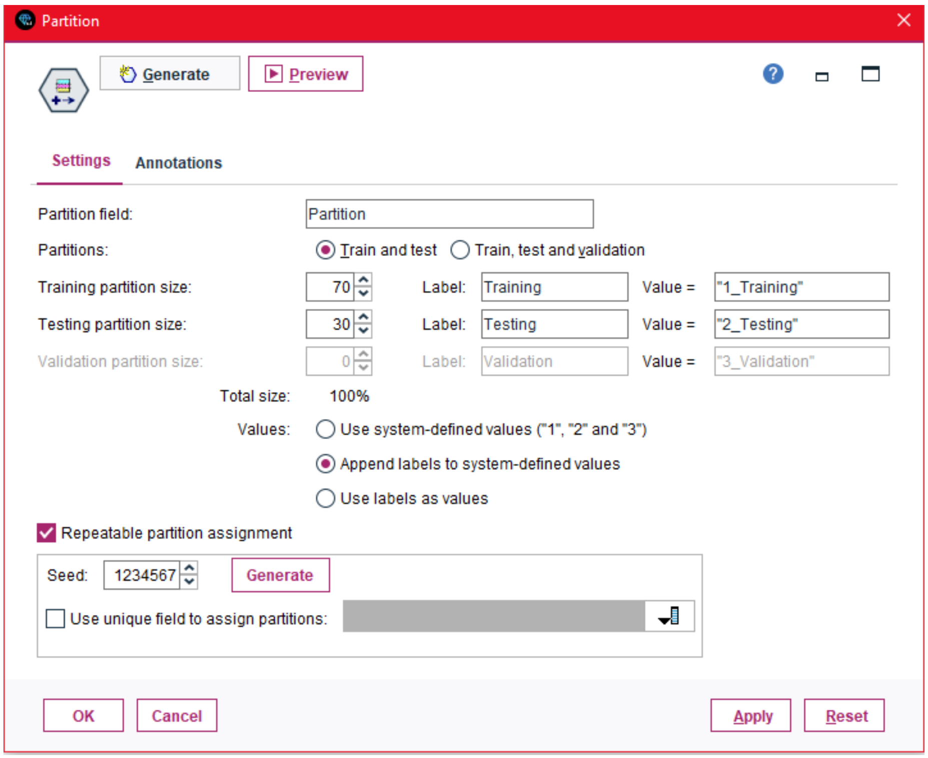
[2] This is a trivial question on a trivial problem: you don't have to be Sherlock Holmes to figure this out :))

PROBLEM 2

In this case you are going to use the full dataset (2201 recs)

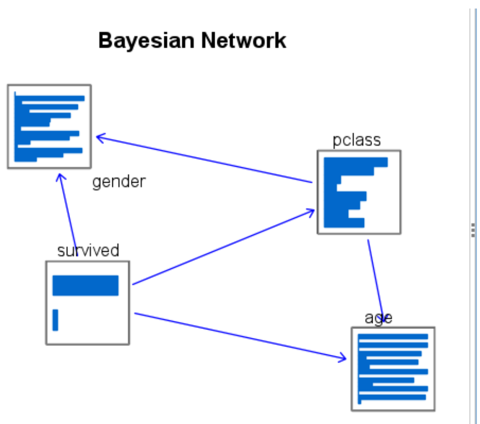
Using Modeler, you must:

a) **Create a TAN classifier, with zero frequency considerations, trained with 70% random data and tested with the other 30%**



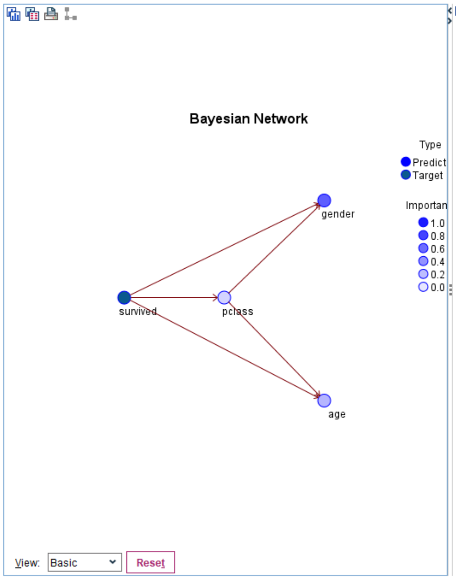
b) **You must report the model parameters (conditional probabilities and prior probabilities for each class) after training the model.**

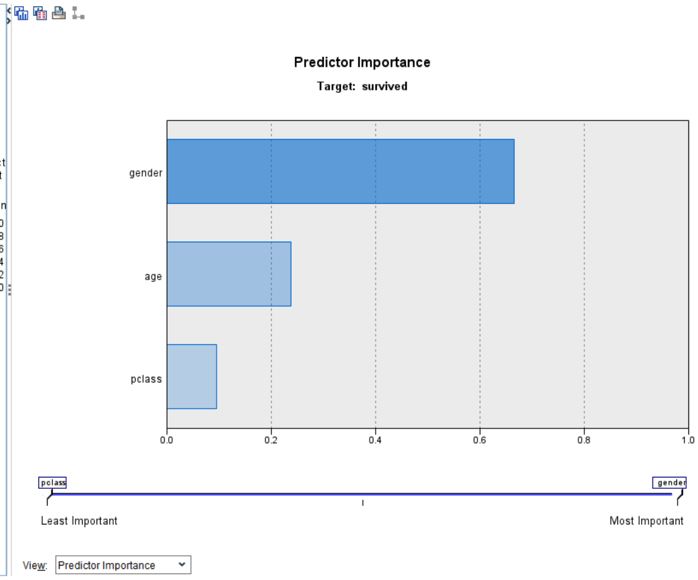
The following figures show the conditional probabilities and prior probabilities after training the model



Conditional Probabilities of pclass

Parents	Probability			
survived	1st	2nd	3rd	crew
yes	0.29	0.17	0.25	0.30
no	0.08	0.11	0.35	0.45





c) **How accurate is the procedure on the training and the test datasets?**

Results for output field survived

Comparing $B- survived with survived

'Partition'	1_Training		2_Testing	
Correct	1,186	78.6%	551	79.62%
Wrong	323	21.4%	141	20.38%
Total	1,509		692	

Coincidence Matrix for $B- survived (rows show actuals)

'Partition' = 1_Training	no	yes
no	991	29
yes	294	195
'Partition' = 2_Testing	no	yes
no	462	8
yes	133	89

Performance Evaluation

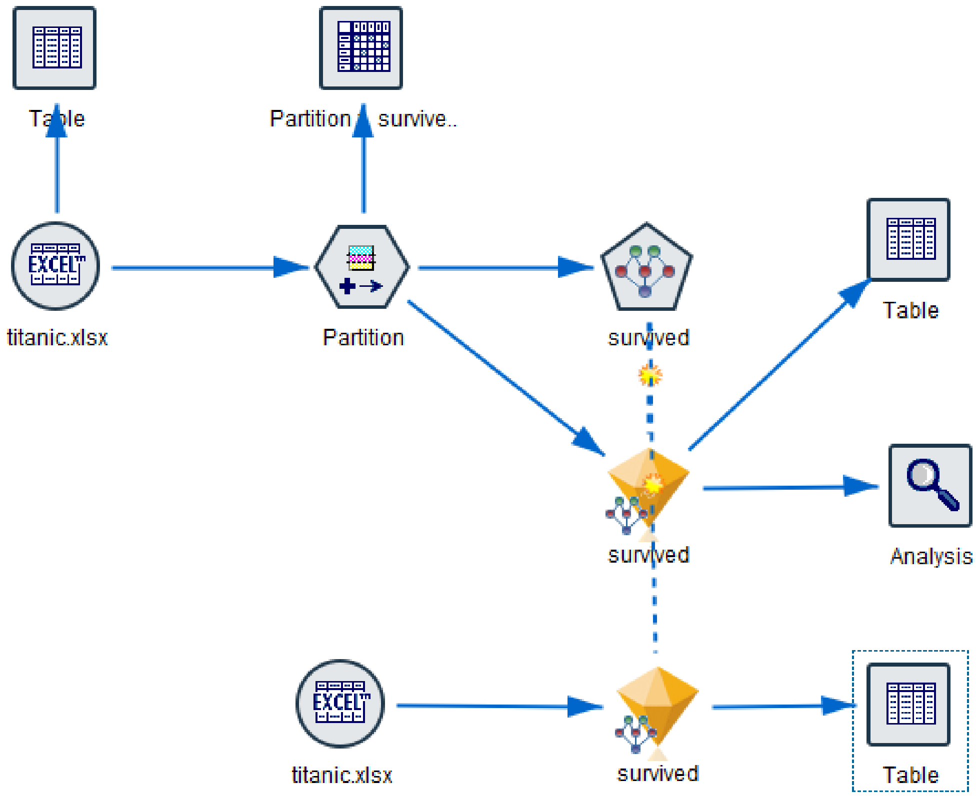
'Partition' = 1_Training	
no	0.132
yes	0.988
'Partition' = 2_Testing	
no	0.134
yes	1.051

The accuracy is 78.6% for training data and 79.62% for testing data. Since both are nearer, the model can be considered accurate.

d) **Once again, suppose that you look at a given individual record: (crew, adult, male). What would your TAN model predict about the fate of this individual[3]?**

When the model is built as shown in below,

[3] This is a trivial question on a trivial problem: you don't have to be Sherlock Holmes to figure this out :))



The results for scoring is as shown below, which says that for the given record, there is chance of not surviving with a probability of 77.7%

	pclass	age	gender	survived	$B- survived	$BP- survived
1	crew	adult	male	$null$	no	0.777