DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

EXAM ASSIGNMENT

Study Programme and level	**MSc Business Intelligence**				
Term	S16r				
Course name and exam code(s)	Bayesian Networks				460152E040
Exam form and duration	Written reexam, WOAI				3 hours
Date and time	10 August 2016				14.00-17.00
Supplementary material allowed	PC necessary, open book, internet allowed				
Other relevant information	A dataset is uploaded 24h before the exam on Blackboard. The students were informed to download them and bring them at the exam together with their computer.				
Hand-in of hand-written material allowed	Yes		No	X	Comments:
Number of pages (incl. front page)	3				

Practical Information
Your exam paper must comply with the following format requirements:

- Your **student ID number** must appear on every page.
- Write **page numbers and total number of pages** on all pages of your paper (e.g. 1 of 15, 2 of 15 and so on).
- Your exam paper MUST be handed in as one **PDF** file, but additional material/appendices may be uploaded in other file formats.
- **The file name** must be your student ID number **AND** the name of the exam.

Exercise:

The following datasets were extracted and preprocessed from the USA census bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html (Donor: Ronny Kohavi and Barry Becker, Data Mining and Visualization Silicon Graphics). The objective is to design an accurate classifier to be able to determine, whether an individual earns over $50K (i.e. $50.000) per year and the demographic variables that are most likely to explain it.

The datasets are:

Data.csv - to build and train the model (n = 45,222 records).
Datatest.csv - to test the model (n = 3,504 records).

Each record consists of 14 attributes, containing:

- Socio-demographic features
- The target variable (attribute 14) with two categories: >50K and <=50K

Requirements:

- Learn different Bayesian Networks and evaluate their capability of predicting, which persons are more likely to earn over $50K/year. Evaluate the networks predictive power using different performance measures. (Note: ROC/AUC from k-fold cross-validation is not required).
- Describe, which are the variables that best explain whether or not a person makes over 50K a year. The descriptions and accompanying interpretation must be comprehensible for somebody with no prior knowledge of BNs.
- Give an example of prediction and diagnostic inference using your network.

Note: All the analyses should be done in *R* statistical software.

Attribute Information:

- V1: age: originally continuous. It has been discretized into 3 intervals.

- V2: workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

  Note:
  Self-emp-inc: refers to people who work for themselves in corporate entities.
  Self-emp-not-inc: refers to people who work for themselves in other legal entities.

- V3: finalweight: originally continuous (it is an internal weight given to individuals based on five particular demographic characteristics: US state, Hispanic origin, race, age and sex; people with similar demographic characteristics should have similar weights). The original variable has been discretized into 3 intervals.

- V4: education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

- V5: education-number (years): originally continuous. It has been discretized into 3 intervals.

- V6: marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

  Note:

  Married-civ-spouse: refers to married with a civilian spouse
  Married-spouse-absent: refers to married but the spouse has a different place of residence for any other reason except separation
  Married-AF-spouse: refers to married with a spouse serving away from home in the Armed Forces

- V7: occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

- V8: race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

- V9: sex: Female, Male.

- V10: capital-gain: originally continuous. It has been discretized into 3 intervals.

- V11: capital-loss: originally continuous. It has been discretized into 3 intervals.

- V12: working hours-per-week: originally continuous. It has been discretized into 3 intervals.

- V13: native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad & Tobago, Peru, Hong, Holland-Netherlands.

- V14: >50K, <=50K. (Target variable)