First name	Name	Matr.-Nr.

RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. van der Aalst

Trial Exam
Introduction to Data Science
January 25th, 2019

Study course:

- Diplom Informatik
- Master SSE
- Master Informatik
- Master Media Informatics
- Master Data Science
- Other: ____________________

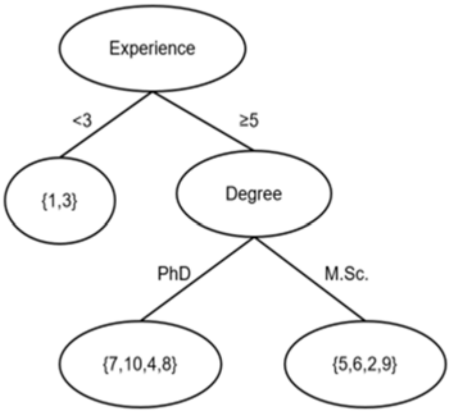
- Duration of the exam:
- Write your first name, name and Matrikelnummer on each sheet.
- Give your solutions in readable and traceable manner. **Solutions will be graded based on completeness and correctness of description/application of the algorithm/method**.
- Give your solutions on the exam sheets only. If you need extra paper, use only paper provided by the supervisors. Make sure to specify your name and Matrikelnummer on all papers.
- Please cross out those things you do not wish to be graded.
- In case of attempted deception, your exam will be graded as **failed**.
- At the end of the exam hand in your complete copy. Do not separate any sheets by removing the staples.
- **You may only use a black or blue pen, and a basic calculator; no additional material (e.g. books, cell phones, laptop, etc.) is allowed.**
- **Only answers that are given in English will be graded.**

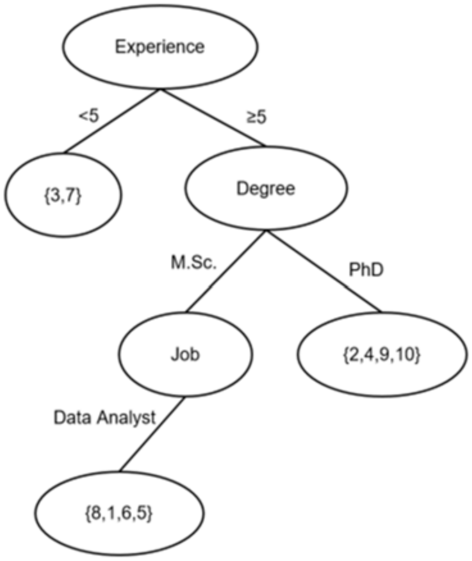
Signature ___________________________

First name	Name	Matr.-Nr.

1. Decision Tree

Suppose that we have two different decision trees to classify people with respect to their salaries (leaves of the trees). Which classifier is better when the impurity is considered as the variance within a leaf?





First name	Name	Matr.-Nr.

2. Regression

Television time / day in hours	Length of deep sleep/day in h
0.3	5.8
2.2	4.4
0.5	6.5
1.8	5.0
0.2	6

a. Based on the data minimize the squared error function to calculate the linear regression function that predicts the length of deep sleep based on the television time.

b. Interpret the regression function obtained in question a). What is the relation between television time and the length of deep sleep described by the function?

First name	Name	Matr.-Nr.

3. Neural Networks

Draw a neuron or a neural network for the Boolean function ((a AND b AND c) OR d). Is it possible to have the function with one neuron? Specify the weights of the network. (Activation function is step function with threshold of zero (0))

First name	Name	Matr.-Nr.

4. Evaluation of Supervised Learning Algorithms

The following table summarizes the results of a statistical regression. Compute the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics.

Forecast	**Observation**
7.1	**6.3**
10.9	**10.2**
12.3	**14.0**
10.5	**10.9**
10.3	**7.7**
8.8	**5.2**

First name	Name	Matr.-Nr.

5. Clustering

Let C_1 and C_2 be two clusters generated in the nth iteration by k-means algorithm, where C_1 contains the elements (11, 68), (37, 52), (28, 70), (37, 55), (44, 56), (22, 57) and (45, 52), C_2 contains the elements (42, 66), (59, 62), (43, 63), (48, 64), (42, 61), (53, 61), (47, 61), (34, 67) and (48, 53). Each element is characterized by two values. When the (n+1)th iteration of k-means algorithm finishes:

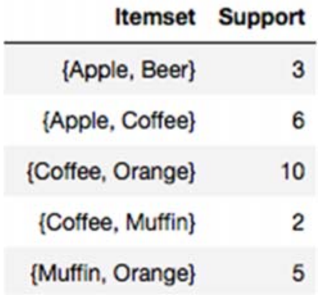
a. Please write down the elements for cluster C_1 and C_2 respectively (use Euclidean Distance).

b. Please write down the centroids for cluster C_1 and C_2 (use Euclidean Distance).
The results should be accurate up to 2 decimal places

First name	Name	Matr.-Nr.

6. Frequent Items Sets and Association Rules

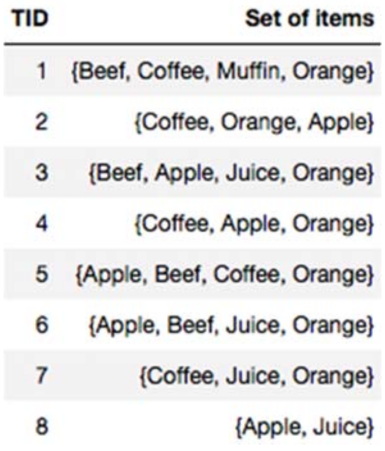
The table below shows a set of frequent itemsets L_2 of length 2. Let C_3 be the set of candidate itemsets of length 3 which is generated from L_2 by using Apriori algorithm.



Itemset	Support
{Apple, Beer}	3
{Apple, Coffee}	6
{Coffee, Orange}	10
{Coffee, Muffin}	2
{Muffin, Orange}	5

a. Please show the two versions of C_3 generated after the itemsets merging step and the pre-pruning step.

Presume that there are three association rules from the transaction set shown in the table below, which are {Beef, Coffee} => {Orange}, {Coffee} => {Apple, Orange} and {Apple} => {Orange}.

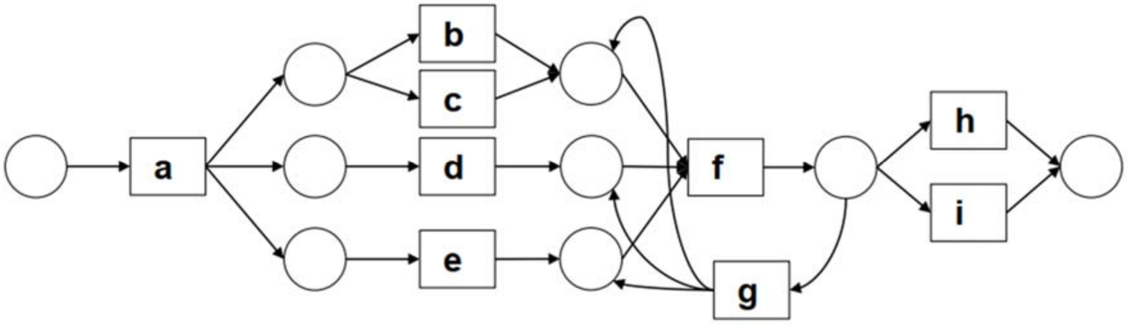


TID	Set of items
1	{Beef, Coffee, Muffin, Orange}
2	{Coffee, Orange, Apple}
3	{Beef, Apple, Juice, Orange}
4	{Coffee, Apple, Orange}
5	{Apple, Beef, Coffee, Orange}
6	{Apple, Beef, Juice, Orange}
7	{Coffee, Juice, Orange}
8	{Apple, Juice}

b. Please calculate and write down the confidence and the lift of each of the three association rules. The results should be accurate up to 3 decimal places.

First name	Name	Matr.-Nr.

7. Process Mining



Compute the fitness of the following event log on the given Petri net using token-based replay. For each trace, indicate the token counts.

[<a, d, c, e, f, g, f, h>^{70},
<a, d, c, f, h>^{20},
<a, d, c, e, f, g, f, h, f>^{10}]

First name	Name	Matr.-Nr.

8. Text Mining

Given the following document corpus, compute the bigram probability estimates for the following bigrams. Make sure to add padding to the text with reserved symbols for the start and end of a document.

D1: “Process Mining is a discipline that perform analysis on process data.”
D2: “Process Mining is a subfield of Data Science.”
D3: “Data Science provides methods and techniques to obtain analysis from data.”
D4: “Process Mining relates to other Process Science disciplines like Business Process Management.”
D5: “Process Mining and Data Mining are related disciplines.”

B1: process mining
B2: data science
B3: <s> process

First name	Name	Matr.-Nr.

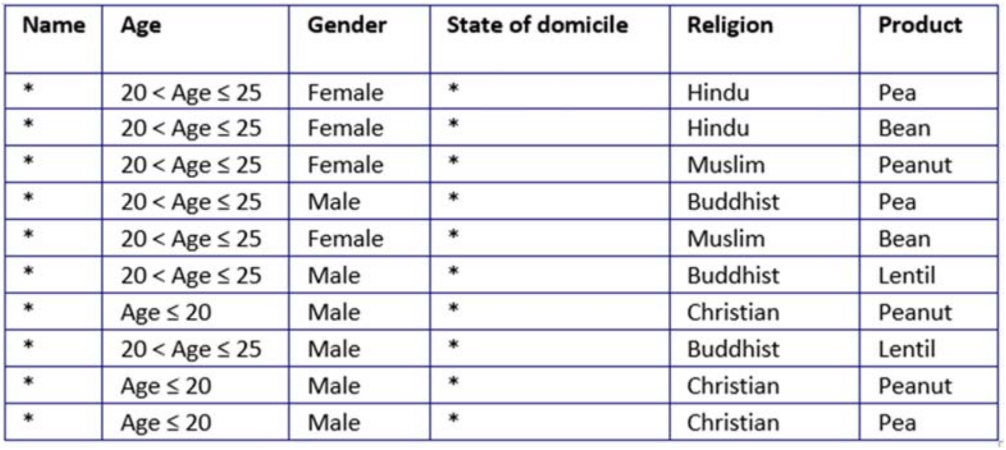
9. Responsible Data Science

For the following data:

a. Given {“Age”, “Gender”, “State of domicile”, “Religion”} as the Quasi-Identifier, does this data have 2-anonymity? If so, identify the equivalence classes.

b. What is the maximum value for l to have distinct l-diversity?

c. What is the maximum value for l to have entropy l-diversity?



Name	Age	Gender	State of domicile	Religion	Product
*	20 < Age ≤ 25	Female	*	Hindu	Pea
*	20 < Age ≤ 25	Female	*	Hindu	Bean
*	20 < Age ≤ 25	Female	*	Muslim	Peanut
*	20 < Age ≤ 25	Male	*	Buddhist	Pea
*	20 < Age ≤ 25	Female	*	Muslim	Bean
*	20 < Age ≤ 25	Male	*	Buddhist	Lentil
*	Age ≤ 20	Male	*	Christian	Peanut
*	20 < Age ≤ 25	Male	*	Buddhist	Lentil
*	Age ≤ 20	Male	*	Christian	Peanut
*	Age ≤ 20	Male	*	Christian	Pea