 Name _____ ID _____

Exercise 1 (22 points) – individual work

- The answers can be typed or handwritten (handwriting must be clear and readable), in this exercise sheet or your own sheet (put your name & ID at the top of the sheet). All answers must be saved to only 1 PDF file.
- Some questions also require the submission of processes/workflows (file.rmp or file.ipynb).
- In case of re-submission (after first grading) or submission after solution is given, your points will be weighted by 0.5.

---

1. (Total 15 points) Retrieve **toydata**. Perform the following tasks to handle missing values.

1.1 (3 points) Consider attribute age

Instructions	Questions
Step 1. Find mean & SD of age, calculated from the whole dataset.	Answer in 2 decimal places Mean of age = SD of age =
Step 2. Identify records with missing age. Impute missing age in these records by central tendency of the whole dataset.	Imputed value =
Step 3. Find mean & SD of age, calculated from the whole dataset, after imputation.	Answer in 2 decimal places Mean of age = SD of age =

1.2 (5 points) Consider attribute gender

Instructions	Questions
Step 1. Find mean & SD of weight and height, calculated from only male records.	Answer in 2 decimal places Mean of weight = SD of weight = Mean of height = SD of height =
Step 2. Identify records with missing gender. Impute missing gender in these records by central tendency of the whole dataset.	Imputed value =
Step 3. Find mean & SD of weight and height, calculated from only male records, after imputation.	Answer in 2 decimal places Mean of weight = SD of weight = Mean of height = SD of height =

1.3 (4 points) Select only attributes: gender, weight, height. Apply KNN imputation with K=5 (other parameters = default) instead of the imputation in step 2 of 1.2.

Instructions	Questions
Step 2. Identify records with missing gender. Impute missing gender in these records by KNN. This time these records may get different imputed values.	Show record ID, imputed gender, weight, height of these records (there are 4 records).
Step 3. Find mean & SD of weight and height, calculated from only male records, after imputation.	Answer in 2 decimal places Mean of weight = SD of weight = Mean of height = SD of height =

1.4 (3 points)

Instructions	Questions
Step 1. Perform age imputation as in 1.1, then select only male records	
Step 2. Identify male records with missing systolic. Apply linear regression imputation using only values from male records (parameters can be default).	Show record ID, age, weight, height, imputed systolic of these records (there are 2 records).
Also submit your workflow that performs both steps. Name the workflow **question1_4.rmp**	

2. (Total 7 points) Retrieve **diabetes**. Perform the following tasks for attribute selection.

Attribute	Short description
preg	Number of times pregnant
plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
pres	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
insu	2-Hour serum insulin (mu U/ml)
mass	Body mass index (weight in kg/(height in m)^2)
pedi	Diabetes pedigree function, i.e. likelihood based on family history
age	Age (years)
Class	tested_positive, tested_negative

 Name _____ ID _____

2.1 (3 points) Apply 3 attribute ranking methods. List 2 most important attributes given by each method in the table below. Note that importance is determined from the magnitude of weight, not just weight (e.g. attribute with weight -0.9 is more important than attribute with weight 0.1)

Method	2 most important attributes
1.	
2.	
3.	

2.2 (2 points) Use Optimize Selection method to obtain an optimal subset of attributes. You can set the workflow in the same way as Forward Selection in the chapter example.

Instructions	Questions
Step 1. Run Optimize Selection	List all selected attributes
Step 2. Compare result from step 1 with results from your attribute ranking methods in 2.1.	The most important attribute for class prediction =

2.3 (2 points) Also submit your workflow that performs 2.1 and 2.2 (they can be put in separate subprocesses). Name the workflow **question2.rmp**.