 Name _____ ID _____

Exercise 5 (14 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 6 points)

Retrieve **cpu data**, which contains the following attributes. All of them are numeric.

Attribute	Description
MYCT	Machine cycle time (nanoseconds)
MMIN	Minimum main memory (kilobytes)
MMAX	Maximum main memory (kilobytes)
CACH	Cache memory (kilobytes)
CHMIN	Minimum channels (units)
CHMAX	Maximum channels (units)
PERF (target)	Performance score

1.1 (3 points) Predict PERF by linear regression method.

Instructions/Questions	Answers
Step 1. Run linear regression to predict PERF in split validation with split ratio = 0.7. When setting linear regression parameters, don't use feature selection & colinear feature elimination.	
Step 2. Run linear regression as in step 1. This time, use feature selection (pick 1 method that gives best result) & colinear feature elimination.	
Step 3. Compare results from steps 1 and 2. Identify attribute(s) that can be considered redundant or can be excluded from the regression model.	Redundant attributes =
Also submit your workflow that performs both steps. The output in each step must include linear regression table and performance vector (RMSE, RRSE, r, and r^2). Name the workflow **question1_1.rmp**.	

 Name _____ ID _____

1.2 (3 points) Use 1 machine learning method (decision tree, SVR, neural network) to predict PERF instead. Try to get better performance than linear regression. If any data preprocessing is needed, do it.

Instructions/Questions	Answers
Submit your workflow. Name the workflow **question1_2.rmp**. Make sure that it output the same performance vector as in 1.1.	Your method =
Briefly explain which data preprocessing is applied to which attribute (if there is any).	Explain preprocessing
Briefly explain how the method's parameters are set differently from classification task.	Explain setting (diff from classification)

2. (Total 8 points) Use training data in the following table to answer questions

2.1 (3 points) Consider AdaBoost model trained by the following training data.

Training Data				
Record	A	B	C	Buy (class)
1	T	T	positive	yes
2	F	T	negative	no
3	F	F	neutral	yes
4	T	F	negative	no
5	F	T	positive	no
6	T	T	positive	yes
7	F	T	negative	no
8	T	F	positive	yes
9	F	F	negative	no
10	T	T	neutral	yes

AdaBoost model
Base classifier 1, weight = 2.197 if C = negative then buy = no if C = neutral then buy = yes if C = positive then buy = yes
Base classifier 2, weight = 2.079 if A = F && B = F then buy = yes if A = F && B = T then buy = no if A = T && B = F then buy = no if A = T && B = T then buy = yes
Base classifier 3, weight = 2.708 if A = F then buy = no if A = T then buy = yes

(a) During 3 rounds of training base classifiers, which training records are easiest to predict?
(b) Find ensemble predictions for <u>record 3</u> and for <u>record 5</u>.

 Name _____ ID _____

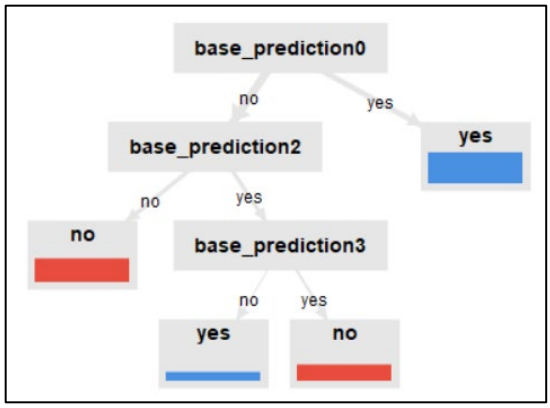
2.2 (2 points) After training 5 weak classifiers, their predictions on training data are as follows.

Row No.	record	buy	base_prediction0	base_prediction1	base_prediction2	base_prediction3	base_prediction4
1	1	yes	yes	yes	yes	yes	no
2	2	no	no	no	no	no	no
3	3	yes	no	no	yes	no	no
4	4	no	no	yes	yes	yes	yes
5	5	no	no	yes	no	no	yes
6	6	yes	yes	yes	yes	yes	no
7	7	no	no	no	no	no	no
8	8	yes	yes	yes	yes	no	yes
9	9	no	no	no	yes	yes	no
10	10	yes	yes	yes	yes	yes	yes

(a) Using voting ensemble by majority vote, find ensemble prediction for record 3.

(b) Find the training accuracy of this voting model (when it predicts all training records).

2.3 (3 points) Suppose that we train meta decision tree classifier to learn the prediction patterns of 5 weak classifiers in 2.2 (i.e. the table in 2.2 is training partition for meta training). The meta model is as follows.



(a) Find ensemble prediction for record 3.

(b) From the meta model, list all base classifiers that contribute to the final prediction.

(c) From the meta model, list all base classifiers that don't contribute to the final prediction