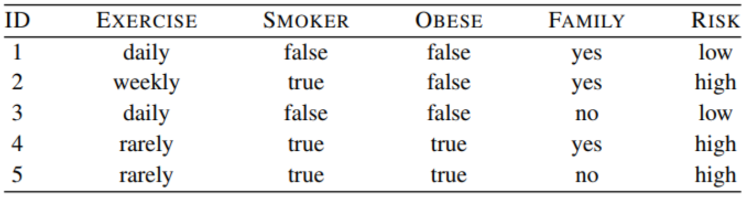
Machine Learning

Tutorial 2

Exercise 1:

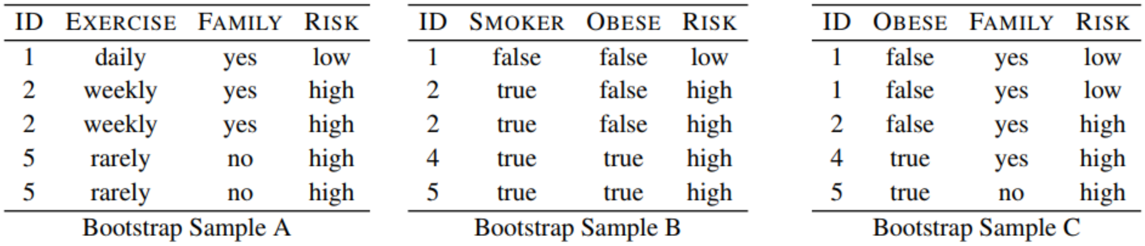
The following table lists a dataset containing the details of five participants in a heart disease study, and a target feature RISK, which describes their risk of heart disease. Each patient is described in terms of four binary descriptive features:

- EXERCISE, how regularly do they exercise.
- SMOKER, do they smoke.
- OBESE, are they overweight.
- FAMILY, did any of their parents or siblings suffer from heart disease.



ID	Exercise	Smoker	Obese	Family	Risk
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

(a) As part of the study, researchers have decided to create a predictive model to screen participants based on their risk of heart disease. You have been asked to implement this screening model using a **random forest**. The three tables below list three bootstrap samples that have been generated from the above dataset. Using these bootstrap samples, create the decision trees that will be in the random forest model (use entropy-based information gain as the feature selection criterion).



ID	Exercise	Family	Risk
1	daily	yes	low
2	weekly	yes	high
2	weekly	yes	high
5	rarely	no	high
5	rarely	no	high

Bootstrap Sample A

ID	Smoker	Obese	Risk
1	false	false	low
2	true	false	high
2	true	false	high
4	true	true	high
5	true	true	high

Bootstrap Sample B

ID	Obese	Family	Risk
1	false	yes	low
1	false	yes	low
2	false	yes	high
4	true	yes	high
5	true	no	high

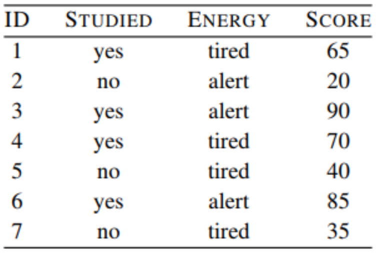
Bootstrap Sample C

(b) Assuming the random forest model you have created uses majority voting, what prediction will it return for the following query:

Exercise=*rarely*, Smoker=*false*, Obese=*true*, Family=*yes*

Exercise 2:

This table lists a dataset of the scores students achieved on an exam described in terms of whether the student studied for the exam (STUDIED) and the energy level of the lecturer when grading the student's exam (ENERGY).

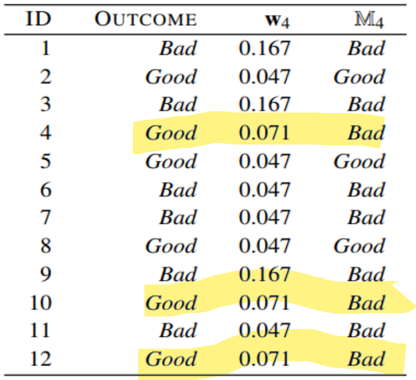


ID	Studied	Energy	Score
1	yes	tired	65
2	no	alert	20
3	yes	alert	90
4	yes	tired	70
5	no	tired	40
6	yes	alert	85
7	no	tired	35

Which of the two descriptive features should we use as the testing criterion at the root node of a decision tree to predict students' scores?

Exercise 3:

The following table shows the target feature, OUTCOME, for a set of instances in a small dataset. An ensemble model is being trained using this dataset using boosting. The table also shows the instance distribution weights, w4, for this dataset used at the fifth iteration of the boosting process. The last column of the table shows the predictions made by the model trained at the fifth iteration of boosting, M4.

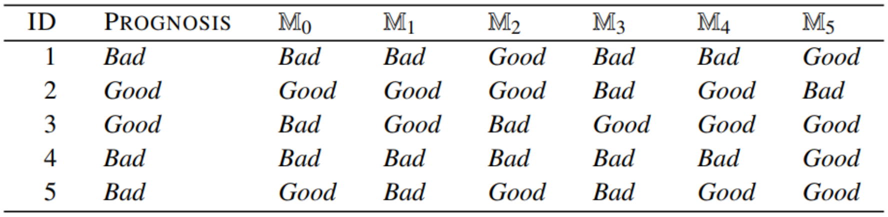


ID	OUTCOME	$\mathbf{w}_4$	$\mathbb{M}_4$
1	*Bad*	0.167	*Bad*
2	*Good*	0.047	*Good*
3	*Bad*	0.167	*Bad*
4	*Good*	0.071	*Bad*
5	*Good*	0.047	*Good*
6	*Bad*	0.047	*Bad*
7	*Bad*	0.047	*Bad*
8	*Good*	0.047	*Good*
9	*Bad*	0.167	*Bad*
10	*Good*	0.071	*Bad*
11	*Bad*	0.047	*Bad*
12	*Good*	0.071	*Bad*

(a) Calculate the error, ϵ, associated with the set of predictions made by the model M4 given in the table above.
(b) Calculate the **confidence factor**, α, associated with M4.
(c) Calculate the updated instance distribution, $\mathbf{w}^{[5]}$, based on the predictions made by M4.

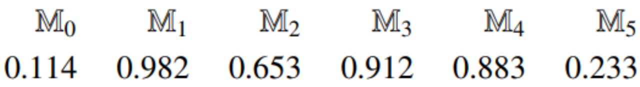
Exercise 4:

The following table shows a set of predictions made by six models in an ensemble and the ground truth of the target feature in a small test dataset, PROGNOSIS.



ID	PROGNOSIS	$\mathbb{M}_0$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$
1	*Bad*	*Bad*	*Bad*	*Good*	*Bad*	*Bad*	*Good*
2	*Good*	*Good*	*Good*	*Good*	*Bad*	*Good*	*Bad*
3	*Good*	*Bad*	*Good*	*Bad*	*Good*	*Good*	*Good*
4	*Bad*	*Bad*	*Bad*	*Bad*	*Bad*	*Bad*	*Good*
5	*Bad*	*Good*	*Bad*	*Good*	*Bad*	*Good*	*Good*

(a) Assuming that these models are part of an ensemble training using **bagging**, calculate the overall output of the ensemble for each instance in the test dataset.
(b) Measure the performance of this bagged ensemble using **misclassification rate** (**misclassification rate** is discussed in detail in Section 9.3; it is simply the percentage of instances in the test dataset that a model has incorrectly classified).
(c) Calculate the overall output of the ensemble for each instance in the test dataset. Assuming that these models are part of an ensemble trained using **boosting** and that the confidence factors, α, for the models are as follows:



$\mathbb{M}_0$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$
0.114	0.982	0.653	0.912	0.883	0.233

(d) Measure the performance of this boosted ensemble using **misclassification rate.**