Practical exercise 2

In this second practice session we will work with our first IA models. This practice session is about decision trees using scikit-learn.

For this practice session we will use a medical insurance dataset. The provided dataset has the following columns.

Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance
- patient: Client identification, is unique for each insurance client
- paid: Boolean indicating if the charge bill is already paid or not.

Answer the following questions.

Model for predicting "**charges**", we want to predict in advance the amount of a bill for a patient.

1) Load the insurance dataset to python. Prepare the dataset for correctly building the models. Justify all the steps and decisions you take. (0.5 point)
2) Build a decision tree for predicting the charge amount of the medical costs. Use all the variables in the dataset, leave all the tree parameters by default. Obtain the R2 and MAPE metrics with the train dataset and the test dataset. Is a good model? Why? Justify your answer (1 points)
3) Delete the "patient" variable and set "max_depth" parameter to 4. Obtain the **R2** and **MAPE** metrics with the train dataset and the test dataset again. Can you see any difference in the scores obtained with the train and test datasets obtained previously? Why is this happening? Is this model better? (1.5 point)
4) Construct the best model you can find for "charges". Justify what you do and why you do each change and step (preprocess, parameters, ...). Correct justifications and procedures will be more valuated than best scores. (2 points)

Model for predicting "**paid**", we want to predict in advance if the bill will be paid or not.

5) Reload the insurance dataset. Build a model for predicting if a bill will be paid or not. Do not delete any variable and leave the default configuration for the model. Obtain the confusion matrix and accuracy of the model for the test dataset. Compute the precision and recall. (1 point)
6) Do the same deleting the variable "patient" and "charges"? Compare the results with the ones obtained before. These are good models? Correct explanations are more important than good codes. (2 point)

7) Construct the best model you can find for predicting "**paid"** variable. Justify what you do and why you do each change and step (preprocess, parameters, ...). Correct justifications and procedures will be more valuated than best scores. Visualize the tree you obtain. (2 points)

Submission. As results will change from one execution to the other due to randomness. You have 14 days to finish and upload the task to Moodle. This practice can be done in pairs or individually. Please submit the python code, commented.