M1 informatique **Machine Learning** Lundi 10 février

Groupe :

Attribute 'Popularity' is numeric. We could still make this into a classification task (how?) but decide to go for regression instead. Does this make sense? In the lectures we were told about linear regression. Would this work here? Why not? Once we are convinced with this we decide to go for regression trees. Let's use the relevant sklearn packages.

**Answer the 4 above questions

**Explain precisely the 2 above lines. Note that if your graphviz doesn't work you can install it with conda install python-graphviz

**Why have we checked if the features were independent? What does "independent" mean? What should we have done if we had found that 2 features were independent?

**We now separate the data into 2 sets. What are these sets X and Y called? Set X will not contain the 'popularity' attribute nor the 'title' attribute. Justify these choices.

**What have we done exactly in the instruction above?

**What has happened? Where is the tree? What do the numbers in the second column mean? Why are there so many null values? Is it not strange that we wanted a maximum depth of 3 but only 3 attributes intervene?

**What does "mse" stand for? Explain now the paradox from the previous question: whay are only 3 attributes used?